


# **Departamento Administrativo Nacional de Estadística**



**Dirección de Regulación, Planeación,  
Estandarización y Normalización  
-DIRPEN-**

**Especificaciones de Imputación  
Encuesta de Consumo Cultural  
-ECC-**

Julio 2008

	<b>ESPECIFICACIONES DE IMPUTACIÓN ENCUESTA CONSUMO CULTURAL -ECC-</b>	CÓDIGO: ME-ECC-EIM-01 VERSIÓN: 02 PÁGINA: 1 FECHA: 08-07-08
ELABORÓ: METODOLOGÍA ESTADÍSTICA	REVISÓ: COORDINADOR DE ESTUDIOS ESTADÍSTICOS	APROBÓ: DIRECTOR DIRPEN

## Especificaciones de imputación

### 1.1) Proceso de Imputación

El proceso de imputación es un mecanismo mediante el cual se corrigen cierto tipo de defectos no deseables en una base de de datos, con opciones sustitutas que, se espera, mejoren la calidad de resultados derivados de dicha base de datos.

Una base de datos, generalmente es una nube de puntos de información de variables codificadas en algún programa especial. Comúnmente, las bases de datos provienen de estudios hechos sobre individuos de interés cuidadosamente seleccionados por algún mecanismo de muestreo, y en otros casos, las bases de datos son consignas de información general de individuos pertenecientes a alguna población específica. De cualquier forma, las bases de datos existen para suplir necesidades de información y para hacer inferencias (cuando se puede) de poblaciones de estudio.

Sin embargo, las bases de datos están sujetas a todo tipo de defectos y fallas operativas, en especial cuando una base de datos no posee información completa de las variables que pretende informar. Esto sucede cuando, al igual que en una investigación, una unidad de estudio o de observación se niega a suministrar la información para alguna variable, y se pierde, o se daña, o deja de existir, o no aplica, o se reconoce que se suministra la información de forma indebida

Las consecuencias de un error o de la falta de información en una base de datos para alguna de sus variables, depende de la forma como los datos hayan sido obtenidos. En caso de que los datos de una base sean censales, la pérdida de información puede no ser tan grave cuando se establecen resultados en niveles de desagregación elevados; sin embargo el dato particular es irrecuperable a menos que se pueda volver a hacer una medición exactamente al mismo individuo o unidad. En otra instancia, si la información que contiene una base de datos es derivada de algún mecanismo muestral, un error o la falta de información en alguna variable, representa multiplicar por miles la carencia dado que el proceso de muestreo le asigna un valor de representación a cada elemento; por tanto, resultados en niveles de desagregación elevados pueden ser erróneos; pero el dato particular es recuperable en cierto grado, aun cuando no haya existido manera de volver a realizar la medición. Es precisamente en este punto donde juegan un papel importante los procesos de imputación

Todo proceso de imputación se emplea para sustituir la no información de una variable en una unidad de estudio por algún dato especial. Es decir, para saber de forma aproximada, cual pudo haber sido la respuesta de una unidad de medición que no suministra la información para alguna variable de estudio de las muchas otras que se le pudo haber medido. Este detalle es importante, porque hay que diferenciar el caso en el que no se puede establecer la información de ninguna de las variables de estudio para un individuo particular y el caso en el que se pierde la información de una o varias variables, pero no de todas, para cierto individuo, es decir, se puede hacer medición sobre la unidad de estudio pero ésta resulta ser no-informante en algunas variables de interés.

A los casos en donde se pudo hacer alguna medición se les denomina como individuos imputables; en los casos donde no se realizó ninguna medición cuando debió haberse hecho se les llama pérdida de muestra o no cobertura. Los problemas de cobertura tienen un tratamiento especial por métodos de muestreo o de corrección censal (sea cual sea el caso), en donde la idea es generar distintos tipos de factores de corrección y de ajuste para compensar las pérdidas, reasignando los factores de expansión mediante fórmulas matemáticas ampliamente justificables.

Por otro lado, para tratar los individuos que se denominan imputables, dependiendo del grado de la pérdida de información, se puede completar la información faltante para cada uno mediante un procedimiento de imputación escogido. Los procedimientos de imputación se seleccionan dependiendo de la naturaleza de los datos, pero principalmente se basan en la cantidad de evidencias observadas para hacer congruencias con individuos similares.

Un procedimiento de imputación consiste en lo siguiente:

- En la base de datos se identifican los individuos que no tienen información en alguna de sus variables.
- Se toma individuo por individuo identificado en el paso anterior y se empieza a filtrar la base de datos por los valores de las variables más importantes y que si tienen información para el individuo a imputar. Se encuentran todos aquellos individuos cuyos valores de las variables coinciden con los valores de las variables del individuo a imputar. Estos individuos se conocen como congruencias o donantes idóneos.
- Se observa el valor de la mediana o moda de las variables que son informadas por los donantes pero no por el individuo a imputar. Este valor es el que imputa la pérdida en el individuo que no informa
- El procedimiento se realiza hasta que todos los individuos tengan información en todas sus variables, es decir, hasta que ya no quede algo más que imputar.

Hay diversos tipos de procedimientos de imputación. Por ejemplo, los comúnmente conocidos como “paquete caliente”, “paquete frío”, de regresión por variables ausentes para encontrar ecuación de imputación, por agrupación de herencia del más semejante, etc. Estos procedimientos consisten, básicamente, en adoptar distintas formas para imputar el valor a partir de la selección de individuos congruentes; por ejemplo, en el método de regresión, en lugar de escoger la mediana de las variables informadas por las congruencias, se ajusta un modelo de tal forma que este mismo aplicado sobre la información que sí posee el individuo a imputar provea del valor no informado.

Diferencias claras entre distintos tipos de métodos de imputación también dependen de las reglas elegidas para imputar. Este es el caso de los paquetes caliente o frío en donde el uno escoge gran número de variables para establecer congruencias, y el otro solo variables importantes de descripción.

También se deben seguir algunas reglas, y lo rigurosas de estas son una distinción para el método de imputación. Tales reglas son:

- El número de individuos que se debe imputar debe ser un cierto porcentaje del total de individuos observados en la base de datos.
- El número de variables no informadas por un individuo debe ser mínimo; y dependiendo del tipo de método elegido para imputar, algunas variables obligatoriamente sí deben ser informadas para llevar a cabo la imputación.
- El tipo de método de imputación se decide con la nube de puntos definitiva, no se debe decidir a priori a la llegada de datos por lo que no puede adivinarse la verdadera naturaleza de cada variable ni calcularse medianas o modas.
- La unidad que se imputa no debe pertenecer a un subconjunto de población extraño o a un dominio poco común.
- En lo posible evitar imputar la información de una variable para un individuo con el promedio (aunque fuere con la información de los individuos congruentes). Es conveniente emplear estadísticos robustos que no se dejen influencias por la presencia de datos atípicos. Recuérdese que la peor imputación es asegurar que la variable toma un valor atípico o poco frecuente.

Finalmente, aunque existe grandes estudios de simulación que aseguran la inclusión de errores relativamente pequeños cuando se imputa el valor de las variables no informadas en lugar de dejarlas NN o NA, junto con una mejor aproximación al cálculo de la desviación, siempre se debe trabajar con la precaución de que los valores imputados pueden generar desvíos en las variables, y se debe informar cuanto es el nivel de imputación en cada variable junto con el procedimiento específico y las limitaciones del mismo.



**ESPECIFICACIONES DE IMPUTACIÓN  
ENCUESTA CONSUMO CULTURAL  
-ECC-**

CÓDIGO: ME-ECC-EIM-01  
VERSIÓN: 02  
PÁGINA: 4  
FECHA: 08-07-08

En particular, para la encuesta de consumo cultural se empleó el procedimiento de imputación de paquete frío, con el cálculo de la mediana de las congruencias en las variables a imputar. Las variables de filtro para seleccionar congruencias fueron las de identificación y aquellas que se consideraron de más importancia temática.



## **BIBLIOGRAFIA**

Fellegi, I. P y D. Holt (1976). "A Systematic Approach to automatic edit and imputation" journal of the American Statistical Association.

Guillermo Ramirez "Imputación de datos" (OCEI-VENEZUELA)

Lohr Sharon L. Muestreo (diseño y análisis). Ed. Thomson

Fernando Medina H. Los métodos de imputación de datos en las encuestas de hogares: teoría y práctica. CEPAL.

Pérez Salvador Blanca Rosa, De los Cobos Silva Sergio. El proceso de depuración de datos, provenientes de una encuesta.

R. Platek. Métodos de Imputación.