DATOS PERDIDOS EN ENCUESTAS DE HOGARES. UN APORTE METODOLOGICO°

Badler ,Clara
Alsina, Sara
Puigsubirá, Cristina
Vitelleschi, María Susana
Arnesi, Nora
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística

1. INTRODUCCION

Las encuestas presentan en general información incompleta que condiciona fuertemente la elección de la metodología a emplear para el análisis y la precisión de los resultados.

Se adopta la definición amplia de encuestas de Lessler y Kalsbeek (3) que denominan así a un estudio científico de una población de unidades tipificadas por personas, instituciones u objetos físicos con el propósito de adquirir conocimientos, observando la población como existe naturalmente y realizando afirmaciones cuantitativas sobre características generales. Dentro de este concepto se incluyen los censos y encuestas a hogares por muestreo probabilístico.

Los resultados del análisis pueden ser establecidos en forma descriptiva o analítica y se denominan errores a aquellos que aparecen cuando existe una discrepancia entre las conclusiones y la realidad.

Una clasificación general divide a estos errores en dos grandes grupos, los de muestreo, que aparecen cuando se trabaja con muestras y no con toda la población, y los ajenos al muestreo, presentes prácticamente en todo tipo de encuesta.

Los esfuerzos para controlar los errores de muestreo están basados en una teoría ampliamente desarrollada, en cambio no sucede lo mismo para evaluar el impacto de los errores ajenos al muestreo. La falta de información conduce al segundo tipo de error.

El tratamiento de este problema debe ser enfocado considerando que se trabaja con una población finita y que en el caso de muestras han sido diseñadas bajo la teoría clásica de aleatorización y las conclusiones estarán basadas en la distribución que determina la selección de la muestra.

A partir de la toma de conocimiento de la existencia de datos perdidos y de sus consecuencias estadísticas, se aborda el problema de su solución, siguiendo los lineamientos propuestos por Little y Rubin (5).

La literatura distingue los métodos para el tratamiento de información incompleta entre aquellos destinados a los casos en que falta la totalidad de la información para algunas unidades del relevamiento y los casos en que falta información sólo para algunas variables.

En este trabajo se enfoca el problema desde la perspectiva de un usuario que cuenta con una base de datos con información incompleta en algunos ítems de ciertas unidades. En diferentes estructuras de pérdidas, generadas mediante métodos de simulación, se evalúa el mecanismo de pérdida mediante la aplicación del test propuesto por Little (4). En un caso específico de falta de información con dificultades de captación se aplican técnicas descriptivas y el algoritmo EM. Si bien el tratamiento se restringe a la etapa de post-relevamiento, es de utilidad en futuros diseños de encuestas para prevenir el problema.

° Financiado por el Programa de Fomento a la Investigación Científica y Tecnológica (resol. C.S. nº202/92) SECYT.

2. CAPTACION DE UNIDADES CON NO RESPUESTA

La falta de información a partir de una base de datos proveniente de una encuesta puede ser captada:

! directamente de blancos o códigos especiales que identifican los datos perdidos.

! indirectamente dado que:

N códigos específicos identifican en forma conjunta a las unidades con no respuesta y a las que no correspondería aplicar la pregunta.

N códigos previstos para alguna categoría de la variable identifican además, a las unidades con no respuesta.

N no se prevé para todas las variables códigos especiales que identifiquen la no respuesta.

Al procesar la información utilizando software estadísticos tales como SAS, SPSS, BMDP, etc., suelen presentarse problemas en las bases de datos disponibles a partir de la categorización y codificación.

Por lo tanto el analista debe realizar la separación de las unidades con datos perdidos a través de su seguimiento, considerando:

! definiciones operacionales de las variables.

! estructura del formulario, teniendo en cuenta el flujo para la obtención de la información correcta.

! instrucciones a encuestadores.

! codificaciones previstas.

! distribuciones básicas de las variables.

3. ESQUEMA DE PERDIDA

Establecido un diseño, las variables son observadas para las unidades seleccionadas y los correspondientes valores se presentan en matrices de datos de unidades por variables.

$$Y_s^* = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & * & \cdots & y_{2p} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ y_{i1} & y_{i2} & \cdots & * \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

Si existe información faltante la matriz original de datos es de la forma: donde * simboliza la información faltante.

Se denomina esquema de pérdida a la estructura que presenta la pérdida de la información en la matriz de datos.

Para una mejor visualización del problema de la falta de información en una muestra seleccionada de una población finita, puede construirse el siguiente esquema:

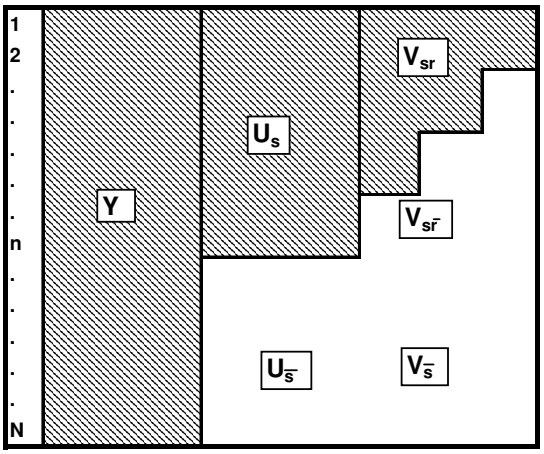

Figura 3.1. Esquema con pérdida de información

Las áreas sombreadas representan los datos. Las unidades son ordenadas en filas: las (N - n) no son muestreadas y de las n muestreadas algunas no son observadas completamente.

Sólo para una mejor comprensión, es útil dividir a las variables "y" en dos submatrices: **U** y **V**, designando con **U** a aquellas para las que se obtuvo información para todos los individuos y con **V** a aquellas con datos perdidos; los valores no muestreados de **U** y **V** se simbolizan con $U_{\bar{s}}$ y $V_{\bar{s}}$, y los muestreados con $\mathbf{U}_s$ y $\mathbf{V}_s$. Esta última a su vez se divide en V_{sr} para los valores registrados y $V_{s\bar{r}}$ para los valores perdidos.

El esquema representa un caso especial de pérdida, monótono o anidado, en el que el conjunto de variables v se presenta o ha sido arreglado en subconjuntos $\mathbf{v}_1,...,\mathbf{v}_k$ tal que $\mathbf{v}_j$ se observa para todas las unidades donde $\mathbf{v}_{j+1}$ también ha sido observado, para j = 1,...,k-1, en otras palabras, $\mathbf{v}_j$ es más observado que $\mathbf{v}_{j+1}$ para todo j.

En general se cuenta con métodos eficientes para manejar valores perdidos con esquema de pérdida monótono, por lo que si no se presenta un esquema de este tipo, en muchos casos puede lograrse despreciando una pequeña cantidad de datos.

Para variables categóricas o categorizadas la información se puede presentar en una tabla de contingencia cuyas celdas están definidas por los niveles conjuntos de las variables.

Al existir valores perdidos algunos de los casos son clasificados completamente y otros sólo parcialmente. Se presenta un caso tridimensional. (Tabla 3.1):

Tabla 3.1. Tabla de contingencia tridimensional con información auxiliar

	y_3				
	PRESENCIA		AUSENCIA		
y_1	y_2		y_2		TOTAL
	PRESENCIA	AUSENCIA	PRESENCIA	AUSENCIA	
PRESENCIA	m_{111}	m_{121}	m_{112}	m_{122}	m_{1++}
AUSENCIA	m_{211}	m_{221}	m_{212}	m_{222}	m_{2++}
	m_{+11}	m_{+21}	m_{+12}	m_{+22}	

Clasificados completamente

	y_3				
	PRESENCIA		AUSENCIA		
y_1	y_2		y_2		TOTAL
	PRESENCIA	AUSENCIA	PRESENCIA	AUSENCIA	
PRESENCIA	r_{111} = ?	r_{121} = ?	r_{112} = ?	r_{122} = ?	r_{1++} = N1
AUSENCIA	r_{211} = ?	r_{221} = ?	r_{212} = ?	r_{222} = ?	r_{2++} = N1
					$r_{..Y1}$

Clasificados parcialmente en y_1

donde $n_{jkl} = m_{jkl} + r_{jkl}$.

Debe distinguirse el problema de datos perdidos de aquél de los ceros estructurales, donde ciertas celdas contienen ceros porque el modelo asigna probabilidad cero a las mismas.

4. MECANISMO DE PERDIDA

Se define como mecanismo de pérdida (proceso de no respuesta) al origen, causas, momento, relaciones, características, que producen la falta de información.

Es importante tratar de establecer si las observaciones han sido perdidas al azar o su falta se asocia a causas definibles. Algunas veces el mecanismo está bajo el control del analista, otras no puede controlarlo pero sí comprenderlo y en muchos casos al no considerarlo explícitamente, se está suponiendo que el mecanismo es ignorable.

Una forma de clasificar el mecanismo de pérdida es de acuerdo a la probabilidad de respuesta:

! Si la misma es independiente de los datos observados y no observados, se dice que el proceso de no respuesta es MCAR, perdidos completamente al azar.

! Si la misma es dependiente de los datos observados se dice que el proceso de no respuesta es MAR, perdidos al azar.

! Si el proceso no es MAR ni MCAR, Molenberghs y Goethebeur (6) lo denominan "informativo".

Es posible formular un modelo estadístico para la distribución conjunta de **Y** y el mecanismo de pérdida:

$f(\mathbf{Y},\mathbf{R} / \Theta, \Phi) = f(\mathbf{Y}/ \Theta)\ f(\mathbf{R} /\mathbf{Y},\Phi)$

donde:

! **R**: es un indicador de datos perdidos.

! $f(\mathbf{Y}/ \Theta)$ es la distribución de densidad conjunta de **Y**.

! $f(\mathbf{R} /\mathbf{Y},\Phi)$ es la distribución del mecanismo de pérdida.

! Θ y Φ parámetros desconocidos a estimar.

En la estimación por verosimilitud, cuando los parámetros que describen el proceso de medición (Θ) son funcionalmente independientes de los parámetros que describen el proceso de pérdida (Φ), es decir en los casos MCAR y MAR, el mecanismo de pérdida es ignorable. En el caso de un proceso "informativo" es no ignorable.

Ante la presencia de un conjunto de datos completos, es posible recurrir a la simulación de las pérdidas para una mejor comprensión de esta clasificación.

5. EVALUACION DEL MECANISMO DE PERDIDA

El acercamiento al conocimiento del mecanismo que condujo a ciertas variables a no ser observadas totalmente es fundamental para:

! la elección del método apropiado de análisis,

! la interpretación de los resultados,

! la precisión de las estimaciones.

La evaluación se puede realizar mediante simples técnicas descriptivas o recurrir a métodos inferenciales.

Test para evaluar MCAR

Para evaluar el supuesto MCAR, cuando las variables son cuantitativas, Little (4) propone un test global basado en la estadística d^2, cuya distribución es sumamente compleja para esquemas generales de datos perdidos, pero se simplifica para algunos, como el monótono.

Si los datos están ordenados en forma monótona, es decir que la variable $\mathbf{y_j}$ se observa más que $\mathbf{y_{j+1}}$ para j=1,...,p-1; n_j es el número de casos para los que $\mathbf{y_j}$ es observado; $n \exists n_1 \exists ... \exists n_p$ y k_j es el número de esquemas con $\mathbf{y_j}$ observado, la estadística resulta:

$$d^2 = SSB_1 / MST_1 + SSB_{2.1} / MST_{2.1} + ...$$

$$+ SSB_{p\text{-}1.12...p\text{-}2} / MST_{p\text{-}1.12...p\text{-}2}$$

donde SSB_1 y MST_1 son, respectivamente, la suma de cuadrados entre grupos y el cuadrado medio total del análisis de la variancia de $\mathbf{y_1}$ basado en los k_1 esquemas de pérdida; $SSB_{2.1}$ y $MST_{2.1}$ son las sumas de

cuadrados entre grupos y el cuadrado medio total del análisis de covariancia de $\mathbf{y}_2$ basado en los k_2 esquemas en los cuales $\mathbf{y}_2$ está presente, ajustado por $\mathbf{y}_1$; los términos restantes se definen en forma similar.

Bajo la hipótesis nula de MCAR y el supuesto de que las variables en estudio se distribuyen normalmente, cada uno de los términos anteriores son independientes, de tal forma que la distribución para muestras pequeñas bajo la hipótesis nula es una suma de funciones de estadísticas F independientes.

$$d^2 = \sum_{j=1}^{p-1} (n_j - 1)(k_j - 1) F_{j.12...j-1} / [n_j - k_j + (k_j - 1) F_{j.12...j-1}]$$

En muestras grandes estas funciones se distribuyen como $\chi 5$ y por lo tanto, d5 tiene una distribución asintótica $\chi 5$ con $\sum_{j=1}^{J} p_j - p$ grados de libertad, siendo J el número de esquemas y p_j el número de variables consideradas en cada uno de ellos.

6. UN METODO DE TRATAMIENTO: ALGORITMO EM

Una vez evaluadas las características del mecanismo que produce la pérdida, se plantea la metodología para intentar solucionar el problema, tratando de utilizar la totalidad de la información disponible.

Basándose en la idea intuitiva clásica de completar los valores perdidos se recurre a un método iterativo ampliamente usado, el algoritmo EM. El mismo formaliza esta idea y permite encontrar los estimadores máximo verosímiles, consistiendo en:

! reemplazar los valores perdidos por los valores estimados;

! estimar los parámetros;

! re-estimar los valores perdidos asumiendo que son correctas las nuevas estimaciones de los parámetros;

! re-estimar los parámetros y así sucesivamente seguir iterando hasta la convergencia.

Cada iteración del algoritmo EM consiste en un paso E y un paso M: el paso M realiza la estimación máximo-verosímil del parámetro θ como si no existieran datos perdidos y el paso E encuentra la esperanza condicional de los datos perdidos dados los datos observados y la estimación de los parámetros, luego los sustituye en los perdidos.

Este algoritmo es fácil de construir, converge confiablemente, pero la convergencia puede ser lenta cuando existe una gran proporción de datos perdidos. No requiere el fuerte supuesto MCAR, sino sólo MAR.

Para el caso de variables categóricas, si se supone que las observaciones de la muestra en estudio provienen de una población con distribución multinomial con parámetros $\boldsymbol{\theta} = (\pi_1, \pi_2, ...\pi_z)$, siendo π_z la probabilidad de clasificar en la celda z, mediante el uso de este algoritmo se distribuyen los datos clasificados parcialmente en la tabla general usando probabilidades condicionales.

$$n_z^{(t)} = E\{ n_z / datos, \pi_1^{(t)}, ..., \pi_Z^{(t)} \}$$

El paso E calcula

$$\pi_z^{(t+1)} = \frac{n_z^{(t+1)}}{n}$$

donde $\{\pi_z^{(t)}, z=1,.....,Z\}$ corresponde a la t-ésima iteración en la estimación de los parámetros y el paso M calcula nuevas iteraciones mediante

La consideración conjunta de las características del esquema que produce las pérdidas y los alcances y limitaciones del método de tratamiento, es el camino que debe seguirse para enfrentar metodológicamente el problema de los datos perdidos.

7. APLICACION DE LA METODOLOGIA

Se trabaja con datos provenientes de encuestas por muestreo probabilístico; los casos presentan características particulares para las variables analizadas y en el origen de la pérdida de datos.

Encuesta de la Juventud en la Ciudad de Rosario

La misma fue realizada por un convenio entre Naciones Unidas y la Escuela de Estadística de la Universidad Nacional de Rosario.

Para visualizar diferentes esquemas de falta de información, se generan distintos porcentajes de pérdidas mediante procedimientos de simulación en la variable número de cuartos ($\mathbf{y_2}$), relevada para unidades de hogares.

Las pérdidas se originan siguiendo dos mecanismos, uno aleatorio y el otro no aleatorio, eliminando el 4%, 8% y 16% de los valores extremos.

Para evaluar la hipótesis que la información faltante está perdida completamente al azar (MCAR) se usa el test propuesto por Little; a tal fin se incluye la variable número de cuartos para dormir ($\mathbf{y_1}$) .(Tabla 7.1)

Tabla 7.1. Evaluación de esquemas de pérdidas

ESQUEMAS			PORCENTAJES DE PERDIDAS		
			4%	8%	16%
ALEATORIO		d^2	0.094	0.511	0.782
		Decisión	no rechazar	no rechazar	no rechazar
NO ALEATORIO	MAXIMO	d^2	30.030	46.553	76.054
		Decisión	rechazar	rechazar	rechazar
	MINIMO	d^2	16.979	26.822	45.505
		Decisión	rechazar	rechazar	rechazar

Se observa que no se rechaza el supuesto MCAR en el primer esquema, en que los valores fueron perdidos en forma aleatoria; en cambio se rechaza cuando los valores perdidos corresponden a los extremos. Esta conclusión es independiente del porcentaje de pérdida. Se corrobora así las características del mecanismo que produjo la pérdida.

Censo Nacional de Población y Vivienda

En encuestas con información incompleta cada variable tiene un tipo de pérdida diferente que condiciona la validez de las estimaciones. La identificación de la falta de información suele presentar difi-

cultades ya que su captación no siempre es directa.

En este contexto, se analiza una de las variables que presenta unidades no observadas por causas no determinadas, de la información censal original para los departamentos Rosario y San Lorenzo, cuestionario ampliado (1980). Al realizar el procesamiento, utilizando SPSS, la categoría "Missing" incluye tanto unidades sobre las que no corresponde relevar la variable, como casos no observados por falta de información debido a otras causas.

Con el objeto de identificar a estas últimas se realiza un monitoreo, teniendo en cuenta definiciones operacionales, diagrama del formulario censal, instrucciones a los censistas, comparaciones cuantitativas y el programa Redatam como auxiliar operativo.

Se trabaja con la variable "hijos tenidos en el último año" (HIJULTAN), de importancia demográfica para la aplicación de las denominadas técnicas de estimación indirecta, que presenta el mayor porcentaje de falta de información debido a otras causas.(Tabla 7.2).

Tabla 7.2. Reclasificación de unidades categorizadas como "Missing" para la variable HIJULTAN

Categorías	Frecuencia Absoluta	% sobre total de la muestra	% excluyendo unid. "no corresp."
No	28083	12.83	52.00
Si, uno o más	4726	2.15	8.74
Missing + No corresp.	164929	75.33	.....
Missing + No resp.	21212	9.69	39.26
Total muestra	218950	100.00	100.00

Se subclasifican los "Missing" en:

! "No Corresponde": personas de sexo masculino, o de sexo femenino fuera del rango de edad de procreación (164 929), ya que la variable HIJULTAN se debe registrar sólo para mujeres entre 14 y 49 años.

! "No Responde": personas que no han respondido por otras causas (21 212), que representan el 39.26 % del total de "reales" respondientes .

Con el objeto de caracterizar las unidades con no respuesta se adiciona información con variables del relevamiento (auxiliares) que no presentan unidades con falta de respuesta y que pueden haber influenciado en la pérdida: "estado civil", "edad", "condición de actividad" y "nivel educacional".

Como forma de evaluar las características del mecanismo de pérdida de la variable HIJULTAN se recurre a métodos descriptivos, observándose en cada cruce los porcentajes de cada categoría de la variable auxiliar sobre el total de casos completos de la variable en estudio y sobre el total de la no respuesta de la misma (Tabla 7.3). Es destacable el diferente comportamiento de la no respuesta en los distintos cruces.

Tabla 7.3. Distribución porcentual de casos completos y no respondientes para la variable

HIJULTAN, según variables auxiliares.

HIJULTAN x Edad		HIJULTAN x Niv.Educ.		HIJULTAN x Est.civil		HIJULTAN x Co-nd.Act.	
CC	No resp.	CC	No resp.	CC	No resp.	CC	No resp.
13.25	67.18	61.21	32.05	79.13	14.27	25.99	47.40
36.00	19.68	28.40	48.00	9.70	2.75	0.26	1.35
34.00	8.58	2.50	6.00	6.82	1.13	2.70	6.61
16.75	4.56	7.89	13.95	4.35	81.89	71.05	44.64

Se continúa el análisis con el esquema HIJULTAN x Estado civil que presenta las características de un esquema Ainformativo≅ .

Al presentarse unidades con información incompleta, las observaciones se esquematizan en dos tablas de contingencia: una completamente clasificada y otra parcialmente clasificada.

Tabla 7.4. HIJULTANxEst.Civil

Tabla de casos completos

	Casado	Unido de hecho	Viudo/Sep./Divorciado	Soltero
No	22344	2457	2147	1135
∃1	3619	728	88	291

Tabla suplementaria de casos clasificados parcialmente por Est.Civil (HIJULTAN faltante)

No Resp.	3028	575	239	17370

A través del algoritmo EM se distribuyen los datos clasificados parcialmente en la tabla asignándolos a las celdas de casos completos. A partir de ello se obtienen los estimadores máximo verosímiles de las frecuencias de las celdas.

Tabla 7.5. Aplicación del algoritmo EM

Probabilidades estimadas Asignación de casos

Paso 1

0.6810	0.0749	0.0654	0.0346
0.1103	0.0222	0.0027	0.0089

24950	2901	2377	15031
4041	859	97	3765

Paso 2

0.4618	0.0537	0.0440	0.2782
0.0748	0.0159	0.0018	0.0697

24948	2900	2377	15031
4043	860	97	3765

Paso 3

0.4618	0.0537	0.0440	0.2782
0.0748	0.0159	0.0018	0.0697

Es de interés comparar los resultados obtenidos, con aquellos que surgen de trabajar sólo con las unidades completas, en cuyo caso se estiman las probabilidades de cada celda a través de las proporciones del primer paso de la aplicación del algoritmo. Las estimaciones máximo verosímiles difieren de las anteriores pues la tabla de casos clasificados parcialmente contribuye a la estimación. A tal fin (Tabla 7.6) se comparan estimadores para :

! proporción de mujeres entre 14 y 49 años que no tuvieron hijos en el último año con respecto al total de este grupo. (π_1)

! proporción de mujeres entre 14 y 49 años que tuvieron uno o más hijos en el último año con respecto al total de este grupo. (π_2)

! proporción de solteras entre 14 y 49 años que tuvieron uno o más hijos en el último año con respecto al total de mujeres entre 14 y 49 años. (π_{24})

Tabla 7.6. Estimaciones calculadas con casos completos (CC) y máxima verosimilitud (MV)

Estimadores	CC	MV
π_1	0.8560	0.8377
π_2	0.1440	0.1623
π_{24}	0.0089	0.0697

Es destacable la diferencia entre los estimadores obtenidos por ambos métodos para la categoría Asoltero≅, resultado previsible por la importancia proporcional de la no respuesta en la misma.

8. CONSIDERACIONES FINALES

Al enfrentarse con datos incompletos en encuestas de hogares es necesario tener en cuenta diferentes aspectos:

! Establecer métodos de captación, rigurosos y objetivos, que permitan discriminar la falta de información

para identificarla fácil y correctamente.

! Evaluar la característica del mecanismo que produce las pérdidas.

! Utilizar un método de tratamiento que permita trabajar con toda la información disponible.

Estas consideraciones, realizadas desde la perspectiva del usuario, deben estar presentes desde el momento de diseñar las encuestas, como una forma tendiente a mejorar la calidad de la información.

9. BIBLIOGRAFIA

1. CENSO NACIONAL DE POBLACION Y VIVIENDA 1980. INDEC. Información Interna. Argentina.

2. DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B..(1977*)."Maximum LIkelihood Estimation from Incomplete Data via the EM Algorithm"*, Journal of the Royal Statistical Society, Serie B, 39.

3. LESSLER, J.; KALSBEEK, W.. (1992). *ANon Sampling Error in Survey≅.* John Wiley & Sons, New York.

4. LITTLE, R. J. A .(1988*). AA Test of Missing Completely at Random for Multivariate Data with Missing Values≅.* Journal of the Royal Statistical Society. Vol. 83, N1 404.

5. LITTLE,R. J. A.; RUBIN, D. B.. (1987*). "Statistical Analysis with Missing Data".* John Wiley & Sons, New York.

6. MOLENBERGHS, A; GOETGHEBEUR, E. (1997*). ASimple Fitting Algorithms for Incomplete Categorical Data≅.* Journal of the Royal Statistical Society . Vol. 59. N1 2. Serie B.

7. RUBIN, D. B.; SCHAFER, J. L.; SCHENKER, N..(1988*). "Imputation Strategies for Missing Values in Post-Enumeration Surveys"*, Survey Methodology, Vol. 14, N1 2.

8. RUBIN, D. B.; STERN, H. S.; VEHOVAR, V..(1995). *"Handling "Don't Know"* Survey Responses: The Case of the Slovenian Plebiscite", Journal of the American Statistical Association, Vol. 90, N1 431.