



Instituto Nacional de Estadística

OPOSICIONES AL CUERPO SUPERIOR DE
ESTADÍSTICOS DEL ESTADO

BOE NÚM. 270, DE 12 DE OCTUBRE DE 2020, PÁG. 87165

**Producción Estadística Oficial:
Principios Básicos del Ciclo de
Producción de Operaciones
Estadísticas**

Grupo de Materias Comunes

Índice general

9	Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico.	1
9.1	Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico	1
9.2	Datos, errores, datos ausentes y controles (edits)	4
9.2.1	Tipos de errores	5
9.2.2	Tipos de datos <i>missing</i>	6
9.2.3	Reglas de depuración	7
9.3	Métodos básicos para la depuración e imputación de datos estadísticos .	9
9.3.1	Depuración durante la fase de recogida de datos	10
9.3.2	Métodos modernos de depuración	11
9.3.3	Métodos de imputación	12
9.4	Estrategia de depuración e imputación	12
9.5	El enfoque de imputación completa.	15
9.6	El enfoque combinado.	16
9.7	El enfoque de reponderación completa.	17
9.8	Imputación por reglas estadísticas	18
9.8.1	Imputación por regresión	20
9.8.2	Imputación por el vecino más cercano	20
9.8.3	Imputación <i>hot deck</i>	21
9.8.4	Grupos de imputación	21
9.8.5	Introducción de un residuo seleccionado aleatoriamente	22
	Bibliografía	24

Tema 9

Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico. Datos, errores, datos ausentes y controles (edits). Métodos básicos para la depuración e imputación de datos estadísticos. Estrategia de depuración e imputación. El enfoque de imputación completa. El enfoque combinado. El enfoque de reponderación completa. Imputación por reglas estadísticas.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

T. de Waal, Pannekoek J. y Scholtus S. (2011). *Handbook of statistical data editing and imputation*. Amsterdam: Wiley

C.-E. Särndal y Lundström S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

9.1 Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico

El objetivo de los Institutos Nacionales de Estadística (INEs) es proporcionar estadísticas de gran calidad sobre muchos aspectos de la sociedad, tan actualizadas y exactas como sea posible. Una de las dificultades que surgen a lo largo del proceso de obtención de las estadísticas es el hecho de que tanto las encuestas tradicionales como los datos administrativos que se usan contienen errores que pueden influir en las estimaciones. Con el fin de evitar sesgos e inconsistencias en la publicación de los datos, el INE realiza un proceso de chequear los datos recogidos y corregirlos en caso de que sea necesario. Este proceso de mejora de la calidad de los datos mediante la detección y corrección de errores comprende una gran variedad de procesos, tanto manuales como automáticos, que se denominan *depuración de datos estadísticos*. La depuración de datos estadísticos se lleva estudiando desde mediados de los años 50 (véase, p.ej., [Nordbotten 1955](#)).

Además de errores en los datos, otro factor que complica el trabajo de los INEs es la existencia de datos *missing* (datos ausentes). Esto se puede considerar como otra forma de datos erróneos, que son fáciles de identificar, pero para los que resulta difícil estimar un buen valor.

Los errores aparecen durante el proceso de medida cuando los valores proporcionados difieren de los valores verdaderos. Esto se puede deber a que los verdaderos valores son desconocidos, difíciles de conseguir. Otro motivo podría ser que las preguntas son mal interpretadas o mal leídas por los informantes. Un ejemplo es el denominado error de medida de unidad que ocurre si el informante proporciona los datos en euros cuando se le pide que los indique en miles de euros. Otro ejemplo es que el informante proporcione sus propios ingresos cuando se piden los ingresos del hogar y el hogar está compuesto por más personas además del informante. En el caso de encuestas económicas, los errores también tienen lugar debido a que las definiciones usadas por los INEs no coinciden con las usadas por el sistema contable de la unidad informante. Puede haber, por ejemplo, diferencias en el periodo de referencia usado por las empresas y el periodo solicitado (el año fiscal frente al año natural es un ejemplo). Después de que los datos han sido recogidos, pasarán por varios procesos, como la codificación, la depuración y la imputación. Los errores que surgen a lo largo de estos otros procesos se conocen como errores de procesamiento. Señalamos que, aunque el propósito de la depuración es corregir los errores, también debe tenerse presente que, como proceso, la depuración también puede introducir errores de forma ocasional. Esta situación no deseable surge si el valor de una variable se modifica porque parece ser erróneo cuando en la realidad es correcto. Los datos *missing* aparecen cuando un informante no sabe la respuesta a una pregunta o se niega a dar la respuesta a una determinada pregunta.

Tradicionalmente, los INEs siempre se han esforzado e invertido muchos recursos en la depuración de los datos, ya que se considera un requisito muy importante para publicar estimaciones acuradas y con calidad. En los procesos tradicionales de procesamiento de una encuesta, la depuración era principalmente interactiva (manual) con la intención de corregir todos los datos en detalle. Los errores detectados o las inconsistencias eran corregidas después de contactar con el informante, lo que implicaba un trabajo que requiere mucho tiempo y trabajo. En este tema se verán métodos más eficientes de depuración.

Se ha admitido desde hace tiempo que no es necesario corregir todos los datos en detalle. Varios estudios (véanse, p.ej., [Granquist 1984](#); [Granquist 1995](#); [Granquist 1997b](#); [Granquist 1997a](#); [Granquist 1997c](#); [Granquist y Kovar 1997](#)) han mostrado que, en general, no es necesario eliminar todos los errores de un conjunto de datos para obtener valores publicables fiables. Los principales productos estadísticos son tablas que contienen agregados, que a menudo se basan en muestras de la población. Esto implica que pequeños errores en registros individuales son aceptables. En primer lugar, porque estos errores tienden a cancelarse cuando se agregan. En segundo lugar, porque si los datos se obtienen a partir de una muestra, siempre habrá un error de muestreo en los valores publicados, incluso si los datos recogidos son completamente correctos. Con el fin de

obtener datos de suficiente calidad, normalmente es suficiente con eliminar sólo los errores más influyentes.

Se emplea demasiado esfuerzo corrigiendo errores que no tienen un impacto notable en los valores publicados. Esto se denomina 'sobredpuración'. La sobredpuración no sólo implica un coste, sino que conlleva una gran cantidad de tiempo, que hace que el periodo entre la recogida de datos y la publicación sea innecesariamente largo. En ocasiones, la sobredpuración se llega a convertir en una 'depuración creativa', que incluso resulta negativa para la calidad de los datos, ya que se modifican datos que son correctos. Para más información sobre los riesgos de la sobredpuración y la depuración creativa¹ véanse [Granquist 1995](#); [Granquist 1997c](#); [Granquist y Kovar 1997](#).

Se ha sostenido que el papel de los INEs en la depuración no debería reducirse a la detección y corrección de errores. [Granquist 1995](#) identifica los siguientes objetivos:

1. Identificar la fuente de errores para proporcionar *feedback* sobre el proceso de producción completo.
2. Proporcionar información sobre la calidad de los datos iniciales y finales ².
3. Identificar y tratar los errores influyentes y los *outliers* en los datos individuales.
4. Cuando sea necesario, proporcionar datos individuales completos y consistentes.

Los datos *missing* constituyen un problema bien conocido al que tienen que enfrentarse todos los organismos que recojan datos sobre personas o empresas. Dependiendo de la legislación existente puede ser más o menos importante en cada país. La solución más común es la imputación, donde los valores de los datos *missing* son estimados. Un problema importante de la imputación es preservar la distribución estadística del conjunto de datos. Éste no es un problema sencillo, especialmente para datos de grandes dimensiones.

En los INEs el problema de la imputación es aún más complicado debido a la existencia de limitaciones en forma de restricciones en los edits, o edits a secas, que los datos tienen que satisfacer. Ejemplos de tales edits son que los beneficios y los costes de una empresa tienen que sumar su cifra de negocios. Los registros que no satisfagan este edit son inconsistentes y por tanto se consideran incorrectos.

¹Para las encuestas con estimaciones basadas en muestreo probabilístico, es fácil aprehender el riesgo de manipulaciones excesivas de los datos, pues se están introduciendo fuentes de variabilidad que no se controlan, introduciendo, por tanto, sesgos desconocidos y aumentando la varianza real de las estimaciones.

²Se denominan datos iniciales a la primera versión proporcionada por los informantes y datos finales a los datos una vez depurados y validados

9.2 Datos, errores, datos ausentes y controles (edits)

Durante el proceso de depuración y de imputación de los datos, los registros erróneos, y los valores erróneos dentro de estos registros, se localizan y se estiman nuevos valores para los valores erróneos y los valores *missing*. La depuración consiste en llevar a cabo dos pasos: primero se localizan los valores incorrectos, a esto se le llama a menudo *localización del error*, y a continuación los valores tienen que ser *imputados*, es decir, se tienen que sustituir por valores mejores, preferiblemente, los correctos.

En principio no es necesario imputar los datos *missing* ni los valores erróneos para obtener estimaciones válidas. En su lugar, se pueden estimar las variables objetivo directamente durante la fase de estimación, sin imputar los datos *missing* ni los erróneos. Sin embargo, este enfoque es en la mayoría de los casos prácticos muy complejo. Mediante una primera imputación de los valores *missing* y los erróneos, se obtiene un conjunto completo de datos. Y a partir de este conjunto completo de datos, se obtienen las estimaciones mediante métodos de estimación estándar. Por tanto, la imputación a menudo se lleva a cabo para simplificar el proceso de estimación.

Las técnicas de depuración e imputación se pueden dividir en dos clases principales, dependiendo del tipo de datos a depurar o imputar: técnicas para datos numéricos y técnicas para datos categóricos (datos entre los cuales no hay una relación de orden, datos agrupados o datos para variables ficticias *-dummy-*). Generalmente, hay diferencias importantes entre las técnicas para estos tipos de datos. Los datos numéricos sobre todo se recogen en encuestas económicas (empresas, establecimientos) mientras que los datos categóricos se recogen en encuestas sociales (personas, hogares, viviendas).

La depuración de encuestas económicas suele ser un problema más complejo que la de la mayoría de las encuestas sociales. Dentro de las encuestas económicas distinguimos entre las encuestas coyunturales, pocas variables y con mucha periodicidad (mensual o trimestral) y las encuestas estructurales, con muchas variables, muchas desagregaciones y periodicidad anual. La principal razón es que en las encuestas económicas estructurales hay muchas más reglas de depuración que en las encuestas sociales y las encuestas económicas estructurales contienen muchos más errores que las sociales.

En los últimos años se ha incrementado el uso de datos administrativos en los INEs. La depuración e imputación de datos administrativos para fines estadísticos tiene determinadas características que no se encuentran en las encuestas muestrales. Por ejemplo, si los datos de varios registros se combinan, además de los errores presentes en los registros individuales, también podemos encontrarnos inconsistencias adicionales entre los datos de los distintos registros debidos a los errores que se producen al cruzar los registros o las divergencias debidas a las definiciones en los metadatos. Véase [A. Wallgren y B. Wallgren 2007](#) para una descripción sobre los métodos para las estadísticas basadas en registros administrativos.

9.2.1 Tipos de errores

Uno de los objetivos más importantes de la depuración de datos es la detección y corrección de errores. Los errores se pueden clasificar de varias formas. Una primera distinción importante se hará entre errores sistemáticos y aleatorios. La segunda será entre errores influyentes y no influyentes. La última será entre *outliers* y no *outliers*.

Definición 1

Errores sistemáticos. Este tipo de errores puede ocurrir cuando un informante malinterpreta o lee incorrectamente una pregunta. Por ejemplo, dar la información financiera en euros en lugar de hacerlo en miles de euros, que era como se requería dicha información. Los errores sistemáticos pueden dar lugar a agregados sesgados. Una vez que se detectan, los errores sistemáticos se pueden corregir fácilmente porque se conoce el mecanismo subyacente. Este mecanismo se puede observar bien a lo largo de toda la historia de un informante o transversalmente a la muestra. Los errores sistemáticos, como los errores en las unidades de medida, se pueden detectar a menudo comparando el valor actual de un informante con el de períodos anteriores (meses, años, trimestres), comparando las respuestas a las variables del cuestionario con los valores de variables de registros, o usando el conocimiento de un experto. Los errores de redondeo se pueden detectar probando si los edits de balance que no se verifican lo hacen con un pequeño cambio en el valor de las variables afectadas.

Errores aleatorios. Los errores aleatorios son debidos al azar, son accidentales. Un ejemplo es un valor observado donde un informante por error tecleó un dígito de más. En la estadística, en general, la esperanza de un error aleatorio es cero. Sin embargo, en nuestro caso, la esperanza de un error aleatorio puede no ser cero. Este es, por ejemplo, el caso del ejemplo anterior.

Los errores aleatorios pueden dar lugar a valores atípicos. En tal caso se pueden detectar usando técnicas de detección de *outliers* o de depuración selectiva. Los errores aleatorios también pueden ser influyentes, en cuyo caso pueden ser detectados con técnicas de depuración selectiva. Si los errores aleatorios no dan lugar a valores atípicos o a errores influyentes se pueden corregir de forma automática.

Errores influyentes. Los errores que tienen una gran influencia en los valores publicables se llaman errores influyentes. Pueden ser detectados con técnicas de depuración selectiva.

El hecho de que un valor tenga una gran influencia en las estimaciones no implica necesariamente que el valor sea erróneo. De hecho, en las encuestas a empresas las observaciones influyentes son bastante comunes ya que variables como la cifra de negocios son a menudo muy asimétricas.

Outliers. Un valor, o un cuestionario, se denomina *outlier* si no se ajusta bien a un modelo considerado para los datos observados. Si un único valor es un *outlier*, se llama *outlier* de una variable. Si el cuestionario en su totalidad, o al menos un subconjunto de varios valores, es un *outlier* cuando los valores se consideran de manera simultánea, se denomina *outlier* multivariante. De nuevo, el simple hecho de que un valor sea un *outlier* no implica necesariamente que este valor contenga un error.

Los *outliers* están relacionados con los valores influyentes. Un valor influyente a menudo es también un *outlier*, y viceversa. Sin embargo, un *outlier* también puede ser un valor no influyente y un valor influyente también puede no ser un *outlier*. Los *outliers* a menudo se detectan durante la macrodepuración.

9.2.2 Tipos de datos *missing*

Los datos *missing* implican una reducción del tamaño efectivo de muestra (que se puede resolver sobremuestreando), y en consecuencia un incremento del error cometido en la estimación (que se debe cuantificar mediante la estimación de dicho error). Un efecto más problemático, que no se puede medir fácilmente, es el sesgo de las estimaciones. Si el mecanismo en la falta de respuesta no depende de datos no observados, la imputación puede dar lugar a estimaciones insesgadas sin la necesidad de hacer ninguna hipótesis. En el caso contrario es necesario hacer nuevas hipótesis para reducir el sesgo mediante la imputación.

Una clasificación de los mecanismos de falta de respuesta que se usa a menudo es: completamente *missing* aleatoriamente (MCAR del inglés *missing completely at random*), *missing* aleatoriamente (MAR del inglés *missing at random*) y *missing* no aleatoriamente (NMAR del inglés *not missing at random*); véanse [Rubin 1987](#), [Schafer 1997](#) y [Little y Rubin 2002](#).

Definición 2

MCAR. La probabilidad de que un valor sea *missing* no depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán ni de los valores de las variables auxiliares: olvido de la respuesta o pérdida de parte de los datos durante su procesamiento. En este caso los datos observados se pueden considerar como un subconjunto aleatorio de los datos completos. Desgraciadamente, el MCAR raramente ocurre en la práctica. Formalmente:

$$P(r_j|y_j, \mathbf{x}, \xi) = P(r_j|\xi). \quad (9.1)$$

donde r_j es el indicador de respuesta de la variable objetivo y_j , donde $r_{ij} = 1$ si el registro i contiene una respuesta para la variable y_j , y $r_{ij} = 0$ en caso contrario, \mathbf{x} es un vector de variables auxiliares que siempre tendremos y ξ es un parámetro del mecanismo de falta de respuesta.

MAR. La probabilidad de que un valor sea *missing* depende de un valor de las variables auxiliares, pero no depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán. Por ejemplo, el mecanismo de falta de respuesta para mayores es distinto del de los jóvenes, pero dentro de cada grupo no depende del valor de la variable objetivo; o en caso de encuestas económicas, las diferencias que se dan entre empresas grandes y pequeñas. Formalmente:

$$P(r_j|y_j, \mathbf{x}, \xi) = P(r_j|\mathbf{x}, \xi). \quad (9.2)$$

En este caso es necesario encontrar los grupos de unidades poblacionales adecuados para pasar del MAR al MCAR dentro de cada grupo.

NMAR. La probabilidad de que un valor sea *missing* depende del(los) valor(es) de la(s) variable(s) objetivo que se imputarán y de los valores de las variables auxiliares: pregunta sobre ingresos, habrá más falta de respuesta en caso de altos ingresos. Formalmente:

$$P(r_j|y_j, \mathbf{x}, \xi) \neq P(r_j|\xi), P(r_j|y_j, \mathbf{x}, \xi) \neq P(r_j|\mathbf{x}, \xi). \quad (9.3)$$

Es el caso más complicado y no se puede usar únicamente los datos observados, sino que hace falta modelizar la dependencia de los mecanismos de falta de respuesta sobre el(los) valor(es) de la(s) variable(s) objetivo.

Otra clasificación de los mecanismos de falta de respuesta relacionados es:

Ignorable. En caso de que sea MAR (o MCAR) y los parámetros a estimar sean 'distintos' del parámetro ξ , es decir, el conocimiento de ξ no ayuda en la estimación de los parámetros de interés.

No ignorable. Si el mecanismo es NMAR o el parámetro ξ no es "distinto" de los parámetros de interés o se dan ambos casos.

9.2.3 Reglas de depuración

Los errores en general se detectan con reglas de depuración o edits. Los edits definen los valores admisibles (o razonables) y las combinaciones de valores de las variables en cada cuestionario. Los errores se detectan verificando si los valores son admisibles de acuerdo con los edits, es decir, comprobando si los edits se verifican o no. Un edit se puede formular como

$$e : x \in S_x,$$

siendo S_x el conjunto de valores admisibles de x . Como veremos a continuación, x se puede referir a una única variable o a varias. Si e es falso, el edit no se cumple mientras que de lo contrario el edit se satisface.

Los edits se pueden clasificar en *duros* o *blandos*. Los edits duros son aquéllos que se deben satisfacer para que un cuestionario sea considerado válido. Por ejemplo, un edit

duro para una encuesta a una empresa específica que la variable *Gastos totales* tiene que ser igual a la suma de las variables *Gastos de personal*, *Gastos de capital*, *Gastos de transporte*, y *Otros gastos*. Los cuestionarios en que no se verifiquen uno o más edits duros son considerados como inconsistentes y se deduce que alguna(s) variable(s) en el mismo es errónea. Los edits blandos se usan para identificar valores dudosos que se sospecha que pueden ser erróneos. Algunos ejemplos son (a) un edit específica que los salarios anuales de los empleados deben de ser inferiores a 10 millones de euros o (b) un edit específica que la cifra de negocios por empleado de una empresa no puede ser mayor que 10 veces el valor del año anterior. Si no se verifica algún edit blando hay que seguir analizando los datos para confirmarlos o rechazarlos.

Veamos a continuación varios ejemplos de clases de edits.

Definición 3

Edits univariantes o restricciones de rango. Un edit que describa los valores admisibles de una única variable se llama edit univariante o restricción de rango. Para variables categóricas una restricción de rango simplemente verifica si los códigos de categoría observados para la variable pertenecen al conjunto especificado de código. El conjunto de valores permitidos S_x es

$$S_x = \{x_1, x_2, \dots, x_C\},$$

y consiste en la enumeración de los C códigos permitidos. Por ejemplo, para la variable S_{sexo} podemos tener $S_x = \{0, 1\}$. Las restricciones de rango para variables continuas se especifican generalmente usando desigualdades. Las más sencillas son las restricciones de valores no negativos, es decir,

$$S_x = \{x | x \geq 0\},$$

Algunos ejemplos son *Edad*, *distintos tipos de costes*, etc. También son comunes restricciones de rango que describen un intervalo como

$$S_x = \{x | i \leq x \leq s\},$$

siendo i el límite inferior y s el superior. Algunos ejemplos son valores admisibles de edad, ingresos u horas trabajadas por semana.

Edits bivariantes. En este caso el conjunto de valores admisibles de una variable x depende del valor de otra variable, que denominaremos y , observada en la misma unidad. El conjunto de valores admisibles es entonces el conjunto de pares admisibles de valores (x, y) . Por ejemplo, si x es *Estado Civil* con valores 0 (nunca casado), 1 (casado) y 2 (previamente casado) e y es *Edad*, podemos tener

$$S_{xy} = \{(x, y) | x = 0 \wedge y < 16\} \cup \{(x, y) | y \geq 16\},$$

equivalente a $S_{xy} = \{(x, y) | x - y > 15\}$.

También podemos encontrarnos con edits de razón que se pueden definir como

$$S_x = \{(x, y) | i \leq \frac{x}{y} \leq s\},$$

Por ejemplo, el cociente entre la cifra de negocios y el número de empleados de una empresa en una determinada rama de la industria.

Edits de balance. Los edits de balance son edits multivariantes que establecen que los valores admisibles de un número de variables están relacionadas con una ecuación lineal. Dos ejemplos son:

$$\begin{aligned} \text{Beneficios} &= \text{Cifra de negocios} - \text{Costes totales} \\ \text{Costes totales} &= \text{Gastos de personal} + \text{Otros costes} \end{aligned} \quad (9.4)$$

Los edits de balance son de gran importancia en la depuración de las encuestas económicas.

Como los edits de balance describen relaciones entre muchas variables se consideran edits multivariantes y deberían tratarse como un sistema de ecuaciones lineales. Es conveniente expresar dicho sistema con notación matricial. Si denotamos las variables de las restricciones (9.4) por x_1 (Beneficios), x_2 (Cifra de negocios), x_3 (Costes totales), x_4 (Gastos de personal) y x_5 (Otros costes), el sistema se puede escribir como

$$\begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

o como $\mathbf{Ax}=\mathbf{0}$. Los valores admisibles de un vector \mathbf{x} sujeto al sistema de edits de balance, definido por la matriz de restricciones \mathbf{A} , se puede escribir como

$$S_{\mathbf{x}} = \{\mathbf{x} | \mathbf{Ax}=\mathbf{0}\}. \quad (9.5)$$

9.3 Métodos básicos para la depuración e imputación de datos estadísticos

Antes de explicar los métodos veamos por qué se han desarrollado estos métodos. Los ordenadores se han usado en el proceso de depuración desde hace muchos años (véase p.ej. [Nordbotten 1963](#)). En los primeros años, sin embargo, su papel se limitaba

a comprobar qué edits no se verificaban. Se grababan los datos en una base de datos, el ordenador comprobaba si los datos verificaban los edits especificados y para cada registro se listaban todos los edits que no se verificaban para corregir estos datos. Es decir, se subsanaban todos los cuestionarios en papel que no verificaban todos los edits. Este proceso iterativo continuaba hasta que (casi) todos los registros verificaban todos los edits.

El principal problema de este enfoque es que durante el proceso de depuración manual no se verificaba la consistencia de los registros. El resultado es que un registro que estaba 'correcto' podía incumplir uno o más edits especificados. Dicho cuestionario, por consiguiente, precisaba más corrección. No era excepcional que algunos registros tuvieran que ser corregidos varias veces. Por tanto, no es de extrañar que depurar de esta forma fuera muy costoso, tanto en términos de dinero como de tiempo, estimándose que entre un 25 % y un 40 % del presupuesto total se empleaba en la depuración (véanse p.ej. [Statistical Methodology 1990](#); [Granquist y Kovar 1997](#)).

9.3.1 Depuración durante la fase de recogida de datos

La técnica de depuración más eficiente de todas es no depurar, sino asegurarse de que los datos que se obtienen durante la fase de recogida son los correctos. Si el objetivo es recoger los datos correctos durante la recogida de los mismos, normalmente se usa un ordenador para grabar los datos. Cuando se da una respuesta inválida a una pregunta o existe una inconsistencia entre dos o más respuestas y la recogida se realiza usando un método asistido por ordenador (CAPI, CATI, CASI o CAWI) los errores pueden ser notificados de manera inmediata. (Para más información sobre datos recogidos por ordenador véase p.ej. [Couper y col. 1998](#)). De esta forma estos errores se pueden solucionar preguntando a los informantes de nuevo. Para CASI y CAWI normalmente no se programan todos los edits, ya que el informante se puede sentir molesto y puede negarse a completar el cuestionario cuando los edits saltan a medida que responde el cuestionario indicando que sus respuestas son inconsistentes.

Una vez terminada la fase de recogida por CAPI, CATI, CASI o CAWI estos cuestionarios contienen menos errores que los recogidos mediante cuestionarios en papel ya que los errores aleatorios que afectan a los cuestionarios en papel no pueden ser detectados y corregidos durante la recogida. Además, si se recogen por CASI o CAWI se pueden evitar los edits de balance calculando de manera automática los totales a partir de las partes. Aunque hay evidencias de que los informantes pueden ser menos acurados cuando rellenan un cuestionario electrónico si los totales se calculan de manera automática. Los INEs en los últimos años se han movido hacia el uso de recogida de datos usando *mixed-modes* donde los datos se recogen usando una mezcla de varios métodos de recogida de datos (véanse p.ej. [Leeuw 2005](#)). Esto, obviamente, tiene consecuencias para la depuración de datos.

Un primer inconveniente es que puede resultar costoso a corto plazo, pero no a largo plazo. Otro inconveniente es que los informantes tienen que ser capaces de responder

durante la entrevista, lo que en el caso de las encuestas económicas no es sencillo. En el caso de CAWI esto se puede solucionar fácilmente si se permite responder a la encuesta en varias etapas.

9.3.2 Métodos modernos de depuración

Depuración interactiva. El conocimiento de los expertos se debe utilizar en la medida de lo posible desarrollando herramientas de depuración interactiva efectivas que permitan comprobar los edits específicos durante la recogida o una vez terminada, y, en caso de que sea necesario, corregir los datos erróneos de manera inmediata. Esto es lo que se denomina depuración interactiva o asistida por ordenador. Para corregir los datos erróneos se pueden seguir varios métodos: el informante puede ser contactado de nuevo, los datos se pueden comparar con los de años/meses previos, los datos se pueden comparar con datos de informantes similares, o se puede usar el conocimiento del experto. Hoy en día es un método estándar de depuración tanto para datos numéricos como para categóricos. El número de variables, edits y registros puede ser, en principio, alto. Y la calidad de los datos depurados de esta forma se considera alta.

Depuración selectiva. La depuración selectiva es un término general para varios métodos de detección de errores influyentes y *outliers*. Las técnicas de depuración selectiva tienen por objetivo aplicar depuración interactiva a un subconjunto de registros bien elegidos de forma que el tiempo y los recursos limitados disponibles para la depuración interactiva se empleen en esos registros que afectan más a la calidad de las estimaciones finales a publicar. Las técnicas de depuración selectiva intentan conseguir este objetivo dividiendo los datos en dos flujos: los registros del flujo crítico se depuran de manera tradicional interactiva, mientras que los registros no críticos se depurarán de forma automática.

Macrodepuración. Distinguiremos entre dos formas de macrodepuración. La primera forma se llama a veces el método de agregación. Formaliza y sistematiza lo que todos los INEs hacen antes de la publicación: verificar si las cifras que se publicarán parecen razonables. Esto se lleva a cabo comparando las cantidades de las tablas a publicar con las mismas cantidades en publicaciones anteriores o con publicaciones relacionadas. Sólo en el caso de que se observe un valor inusual, se usa un proceso de microdepuración a los registros individuales y a los campos que contribuyen a la cantidad sospechosa. Una segunda forma de macrodepuración es el método de la distribución. Los datos disponibles se usan para caracterizar la distribución de las variables. A continuación, todos los valores individuales se comparan con la distribución. Los registros que contengan valores que se puedan considerar extraños, teniendo en cuenta la distribución, son candidatos para una mayor inspección y posiblemente para depuración. La macrodepuración, en particular, el método de agregación, siempre se ha usado en los INEs.

Depuración automática. Cuando la depuración automática se utiliza, los registros son depurados por un ordenador sin la intervención humana. En este sentido, la depuración automática es lo contrario a la aproximación tradicional en el problema de depuración,

donde cada registro se depura manualmente. En los últimos años esta depuración se ha perfeccionado mucho ya que los ordenadores son más rápidos y los algoritmos se han simplificado y se han vuelto más eficientes.

9.3.3 Métodos de imputación

Se pueden usar dos aproximaciones. La primera es la imputación manual, que consiste en el recontacto con el informante o el conocimiento del experto para obtener una estimación del valor *missing*. La segunda es la imputación automática, que se basa en técnicas de estimación estadística, como los modelos de regresión.

Un modelo de imputación predice un valor *missing* usando una función de variables auxiliares, que se llaman predictores, y que se pueden obtener de la encuesta actual, de otras fuentes como valores de esa variable en períodos anteriores o de registros administrativos. Los modelos de imputación más comunes son variantes de modelos de regresión con parámetros estimados de los datos correctos observados. En el caso de variables categóricas se suelen usar métodos de donantes, que reemplazan los valores *missing* en un registro con los valores de un registro vecino completo y válido. Los donantes se eligen de forma que se parezcan lo máximo posible al registro con los valores *missing*.

9.4 Estrategia de depuración e imputación

La depuración de datos a menudo se realiza como una secuencia de distintos pasos de procesos de detección y/o corrección. Para finalizar este tema veamos una descripción global de una estrategia de depuración. Esta estrategia se representa en la Figura 9.1, que consiste en los siguientes cinco pasos.

1. *Tratamiento y corrección de errores sistemáticos*. Consiste en identificar y corregir los errores sistemáticos que son evidentes y fáciles de tratar con suficiente fiabilidad. Se puede hacer automáticamente con, virtualmente, ningún coste, y por tanto mejorar tanto la eficiencia como la calidad del proceso de depuración.
2. *Microselección*. Selecciona para su tratamiento interactivo registros que contienen errores influyentes que no pueden ser tratados de manera automática con suficiente fiabilidad. Por tanto serán controlados tanto manualmente (por expertos) como automáticamente (con edits especializados y algoritmos de depuración). En este paso los datos se dividen en dos flujos: uno crítico y otro no crítico, usando técnicas de depuración selectiva. Para saber en qué medida un registro puede contener errores influyentes se puede usar una función *score*. Esta función se construye de forma que los registros con *scores* más altos se consideran como los que contienen efectos importantes sobre las estimaciones de los parámetros objetivo. Para ello se establece un umbral y todos los registros con *scores* por encima del umbral se revisan manualmente, mientras que los que estén por debajo se tratan de forma

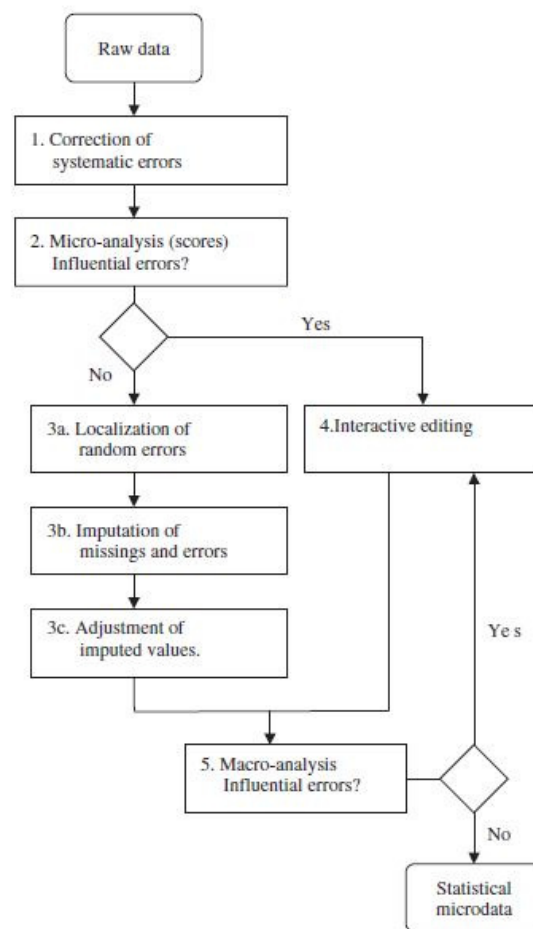


Figura 9.1: Estrategia general de depuración e imputación

automática.

3. *Depuración automática.* Emplea los procedimientos automáticos de detección y corrección automática de errores a los registros que no son seleccionados para la depuración interactiva del paso 2. El primer paso en el tratamiento automático de errores es la localización de errores. Como los errores sistemáticos ya se han eliminado, los errores que todavía existen en este momento son aleatorios. Una vez que los errores duros se han definido y programado es fácil comprobar si los valores de un registro son inconsistentes en el sentido de que algunos de estos edits no se verifican. Sin embargo, no es tan obvio el decidir qué valores son erróneos en un registro inconsistente. A continuación se imputan los datos que faltan de manera automática. El mejor método de imputación para una situación particular dependerá de las características del conjunto de datos y su finalidad. En muchos casos los edits no son tenidos en cuenta por el método de imputación. Como consecuencia, los valores imputados pueden ser inconsistentes con las validaciones. Este problema se puede resolver introduciendo una fase de corrección en la cual se hacen ajustes en los valores imputados de forma que los registros verifiquen los edits y los ajustes sean lo más pequeños posible.

4. *Depuración interactiva*. Se aplica la depuración interactiva a la minoría de registros con errores influyentes. Los errores importantes en empresas grandes que tienen una gran influencia sobre los agregados que se publican y para los cuales no existen modelos de imputación acurados no se consideran adecuados para los procedimientos genéricos de depuración automática. Estos registros son tratados por expertos en un proceso llamado depuración interactiva.
5. *Macrodepuración*. Selecciona registros con errores influyentes usando métodos basados en técnicas de detección de *outliers* y otros procesos que hacen uso de toda o de una gran fracción de las respuestas. Los pasos anteriores usan todos métodos de microdepuración. Estos procesos de microdepuración se pueden realizar desde el principio de la fase de recogida de datos, tan pronto como los registros están disponibles. Por contra, las técnicas de macrodepuración usan información de otros registros y sólo se pueden usar si una gran parte de los datos ya se ha recogido o imputado. Las técnicas de macrodepuración también son técnicas de depuración selectiva en el sentido de que aspiran a prestar atención únicamente a posibles valores erróneos influyentes.

Aunque los procesos automáticos se usan con frecuencia para errores de poca importancia, elegir los métodos más adecuados de detección de errores e imputación es muy importante. Si se usan métodos inapropiados, especialmente para grandes cantidades de errores aleatorios y/o falta de respuesta, se puede introducir sesgo adicional. Más aún, a medida que mejora la calidad de los métodos de localización automática de errores y de imputación, se pueden asignar más registros al tratamiento automático en el paso 3 y menos registros son seleccionados para el paso de depuración interactiva, que resulta mucho más costoso y consume mucho más tiempo.

El flujo de procesos sugerido en la Figura 9.1 es simplemente una posibilidad. Dependiendo del tipo de encuesta, de los recursos disponibles y de la información auxiliar, el flujo de procesos puede ser diferentes. No todos los pasos se realizan siempre y algunos de los pasos puede ser diferente. Para encuestas sociales, por ejemplo, la depuración selectiva no es muy importante porque las contribuciones de los individuos al total publicado no son muy diferentes, al contrario de lo que ocurre con la contribución de las empresas pequeñas y grandes en una encuesta económica. A menudo, en las encuestas sociales, debido a la falta de edits duros, el principal tipo de error detectable es la falta de respuesta.

La UNECE ha desarrollado el Generic Statistical Data Editing Model (GSDEM) (UNECE 2019) como una referencia para todos los estadísticos oficiales entre cuyas actividades se incluya la depuración de datos. El GSDEM incluye las estrategias de depuración e imputación bajo distintos escenarios, encuestas sociales, encuestas económicas (coyunturales y estructurales), censos u operaciones basadas en la integración de datos.

9.5 El enfoque de imputación completa.

Existen muchos métodos de imputación. Cada método tiene variaciones, porque una modificación mínima en un método bien establecido puede ser necesaria para ajustarse a las necesidades de una encuesta particular.

En el enfoque de imputación completa, imputamos todos los valores y_k que son *missing*, independientemente de que haya falta de respuesta parcial o total. El *conjunto de datos completado* resultante es el conjunto de valores $\{y_{\bullet k} : k \in s\}$, donde:

$$y_{\bullet k} = \begin{cases} y_k, & \text{si } k \in r_i ; \\ \hat{y}_k, & \text{si } k \in s - r_i. \end{cases} \quad (9.6)$$

Es decir, $y_{\bullet k}$ toma el valor observado y_k cuando k responde, y el valor imputado \hat{y}_k cuando y_k es *missing*. Todos los valores *missing* han sido reemplazados por sustitutos. El resultado es una matriz rectangular de datos. En este punto, \hat{y}_k denota un valor que podía haber sido construido por cualquiera de los métodos de imputación que mencionaremos.

Los estadísticos descriptivos tradicionales, media, varianza y otros, se pueden calcular a partir del conjunto de datos completo. Por ejemplo, la media del conjunto completo de datos es $\bar{y}_{\bullet k} = \sum_s \frac{y_{\bullet k}}{n}$. Por el contrario, la media que habría sido calculada en el caso de respuesta completa sería $\bar{y}_s = \sum_s \frac{y_k}{n}$. Ambas medias están basadas en valores de n , pero se diferencian en un margen no conocido. De modo similar, nosotros podemos calcular la varianza u otro estadístico común desde el conjunto de datos completo. Estas medias también se diferenciarían en el caso de un hipotético conjunto de datos que consistiera en su totalidad en valores observados.

El objetivo es estimar el total poblacional de la variable de estudio y $Y = \sum_U y_k$. En el cálculo de la estimación, los valores imputados son tratados como datos reales observados, por lo menos cuando se trata de estimación puntual. Uno pretende que los valores imputados sean al menos ‘tan buenos’ como las observaciones verdaderas. Esta perspectiva invita al uso de exactamente el mismo método de estimación que en el caso ideal de respuesta completa. En consecuencia, el elemento k recibe el mismo peso tanto si el valor grabado en el fichero de datos es una observación real y_k o un valor imputado \hat{y}_k . Cabe señalarlo, porque alguien podría argumentar que los valores imputados y los valores realmente observados deberían recibir distinto tratamiento en el proceso de ponderación.

Llamamos *estimador de respuesta completa* al que sería usado si y_k se hubiese observado para todos los elementos de la muestra. Este estimador tiene una fórmula bien definida. Después de la imputación, podemos calcular esta fórmula para el conjunto completado de datos dado en (9.6). El resultado se denomina *estimador imputado*. El sistema de ponderación del estimador de respuesta completa se usa sin modificaciones. El conjunto

de datos completado (9.6) simplemente reemplaza el conjunto de datos deseado (pero no disponible) que consiste en observaciones reales.

Si el estimador de Horwitz-Thompson, $\hat{Y}_{HT} = \sum_s d_k y_k$, se utiliza como el estimador de respuesta completa, entonces el equivalente imputado es

$$\hat{Y}_{IHT} = \sum_s d_k y_{\bullet k} = \sum_{r_i} d_k y_k + \sum_{s-r_i} d_k \hat{y}_k \quad (9.7)$$

Si el estimador de respuesta completa es el estimador GREG $\hat{Y}_{GREG} = \sum_s d_k g_k y_k$, los pesos son $d_k g_k$ para $k \in s$. El *input* es $\sum_U \mathbf{x}_k^*$. El equivalente imputado es

$$\hat{Y}_{IGREG} = \sum_s d_k y_{\bullet k} = \sum_{r_i} d_k g_k y_k + \sum_{s-r_i} d_k \hat{y}_k \quad (9.8)$$

De este modo, en la práctica, la estimación puntual tras la imputación es extremadamente simple, puesto que los pesos no cambian. Sin embargo, la estimación de la varianza es más complicada.

9.6 El enfoque combinado.

El enfoque combinado, ampliamente utilizado, usa la imputación para la falta de respuesta parcial, y la ponderación para compensar la falta de respuesta total. Supongamos que tenemos un vector auxiliar $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$. Con la información correspondiente

$\mathbf{X} = \begin{pmatrix} \mathbf{X}^* \\ \hat{\mathbf{X}}^o \end{pmatrix}$, podemos calcular los pesos calibrados en una etapa (*single-step calibrated weights*) para $k \in r$ con $d_{\alpha k} = d_k$. Son

$$w_k = d_k v_k, v_k = 1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k)^{-1} \mathbf{z}_k. \quad (9.9)$$

Los pesos $d_k v_k$ tienen incorporada una compensación para la falta de respuesta parcial y la propiedad de calibrado deseada $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$. Son apropiados si la variable y se ve afectada sólo por la falta de respuesta total, es decir, cuando $r_i = r$. El estimador calibrado ponderado apropiado para el total $Y = \sum_U y_k$ es entonces

$$\hat{Y}_W = \sum_r d_k v_k y_k. \quad (9.10)$$

Lo más probable es que la variable y se vea afectada tanto por la falta de respuesta parcial como total. Entonces los valores y_k están disponibles sólo para $k \in r_i \subset r \subset s$. El enfoque combinado necesita un primer proceso de imputación seguido por un proceso de ponderación. Primero procedemos a la imputación de los elementos con falta de

repuesta parcial, $k \in r - r_i$, con el fin de crear una matriz de datos rectangular completa con valores específicos para cada variable de estudio y para cada elemento k en el conjunto de respuestas r (mientras que en la imputación completa, definida en (9.6), también imputamos para $k \in s - r$). El conjunto completo de datos para la variable y es $\{y_{\bullet k} : k \in r\}$, donde

$$y_{\bullet k} = \begin{cases} y_k, & \text{si } k \in r_i; \\ \hat{y}_k, & \text{si } k \in r - r_i. \end{cases} \quad (9.11)$$

siendo $y_{\bullet k}$ el valor observado de y_k cuando k responde a este ítem, y en otro caso $y_{\bullet k}$ es el valor imputado \hat{y}_k . El método usado para construir \hat{y}_k puede ser cualquiera de los incluidos en los dos últimos apartados del tema. Estimamos el total $Y = \sum_U y_k$ sumando los valores adecuadamente ponderados $y_{\bullet k}$ sobre el conjunto de respuestas r . El enfoque combinado proporciona el estimador del ítem imputado y calibrado (del inglés *item imputed calibration estimator*).

$$\hat{Y}_{IW} = \sum_r d_k v_k y_{\bullet k} = \sum_{r_i} d_k v_k y_k + \sum_{r-r_i} d_k v_k \hat{y}_k \quad (9.12)$$

donde los pesos calibrados son $d_k v_k$, con v_k definido en (9.9).

9.7 El enfoque de reponderación completa.

Una alternativa al enfoque combinado es una dependencia completa en los pesos. No se usa ninguna imputación. El estimador se obtiene como la suma de respuestas ponderadas apropiadas y_k sobre el conjunto de respuestas r_i . Con este fin usamos los pesos calibrados dados para $k \in r_i$ por

$$w_k = d_k v_{ik}, v_{ik} = 1 + (\mathbf{X} - \sum_{r_i} d_k \mathbf{x}_k)' (\sum_{r_i} d_k \mathbf{z}_k \mathbf{x}_k)^{-1} \mathbf{z}_k. \quad (9.13)$$

Si $r_i = r$ en (9.13), usamos la fórmula (9.9) apropiada para el caso en que sólo hay falta de respuesta total. Los pesos $d_k v_{ik}$ compensan tanto la falta de respuesta total como parcial. Amplían, por así decirlo, del conjunto de falta de repuesta parcial r_i a la muestra s , saltándose r . Verifican la propiedad de calibrado deseada $\sum_{r_i} d_k v_{ik} \mathbf{x}_k = \mathbf{X}$. El estimador de reponderación completa calibrado (del inglés *fully weighted calibration estimator*) para el total $Y = \sum_U y_k$ es entonces

$$\hat{Y}_{FW} = \sum_r d_k v_{ik} y_k. \quad (9.14)$$

con v_{ik} definido en (9.13). Si todos los r_i son distintos, el procedimiento requiere distintos pesos para cada variable de estudio. Esto se puede ver en la práctica como menos atractivo. Una razón es que uno puede no desear cargar el fichero con los datos de la encuesta con tantos conjuntos de pesos como variables de estudio hay en la encuesta.

El estimador del ítem imputado \hat{Y}_{IW} dado por (9.12) y el estimador de reponderación completa \hat{Y}_{FW} dado por (9.14) usan distintos sistemas de pesos (a no ser que $r_i = r$), pero

tienen una cosa en común: los pesos se calculan sobre el mismo conjunto de información auxiliar, $\mathbf{X} = \begin{pmatrix} \mathbf{X}^* \\ \hat{\mathbf{X}}_o \end{pmatrix}$. Los pesos $d_k v_{ik}$ en (9.14) son menos en número, numéricamente mayores en media, y crean un aumento mayor que los pesos $d_k v_k$ en (9.12). Un punto importante es que (9.14) y (9.12) son idénticos para algunos tipos de imputación.

Comentario 1. Otra situación donde un procedimiento completo de pesos calibrados se considera inconveniente es en el de la validación cruzada. Por ejemplo, al considerar una encuesta de salud en la que la variable ‘estado de salud’ (variable categórica A) se cruce con la variable ‘tipo de actividad profesional’ (variable categórica B). Entonces, en el caso de tener un atributo definido por una categoría particular de A y otro de B y deseemos estimar el número de personas en una población con el atributo que cruce ambas variables. La variable de estudio dicotómica y tiene el valor $y_k = 1$ si la persona k tiene el atributo e $y_k = 0$ en caso contrario. El parámetro objetivo es $Y = \sum_U y_k$, el número de elementos de la población con ese atributo.

En el procedimiento totalmente ponderado con pesos definidos por (9.13), el conjunto de respuestas del ítem r_i es el conjunto de elementos que han respondido tanto a A (indicando una de las categorías totalmente exhaustivas de A) como a B (indicando una de las categorías totalmente exhaustivas de B). Otras clasificaciones cruzadas pueden ser de interés en la encuesta, digamos ‘estado de salud’ (variable A) cruzada con ‘grado de actividad física’ (variable C). Para estimar el número de elementos en una celda de esa clasificación cruzada, se debería de identificar y usar un nuevo conjunto r_i como base para calcular los pesos (9.13). En otras palabras, en el enfoque de ponderación completa, cada clasificación cruzada nueva de interés precisa un nuevo conjunto de pesos calibrados. Esto se puede ver como un inconveniente, por eso el enfoque combinado a menudo es preferible.

9.8 Imputación por reglas estadísticas

Veamos a continuación algunas reglas usadas comúnmente para calcular valores imputados \hat{y}_k , haciendo especial énfasis en el enfoque que combina ponderación e imputación.

Es necesario distinguir la imputación por una regla estadística de la imputación especial. Los métodos del primer caso se derivan de argumentos de la predicción estadística. Estos métodos se encargan de la mayor parte de la imputación en una encuesta. La imputación especial, por juicio del experto y por datos históricos, se reserva para unas pocas unidades influyentes.

Las reglas estadísticas de imputación más usadas son: la imputación por regresión, la imputación por el vecino más cercano y la imputación por *hot deck*. Son categorías

amplias, y a continuación se proporcionarán las ideas básicas de cada una de ellas. Casos especiales de la imputación por regresión son imputación por la media e imputación por razón.

La imputación por regresión y por el vecino más cercano precisan información auxiliar. Y dan lugar a imputaciones *determinísticas*. El vector auxiliar utilizado para imputar se denota por $\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})^t$ y está compuesto por los valores de una o más *variables de imputación*. Asumimos que el vector de valores \mathbf{x}_k es conocido para todos los $k \in s$. Cuando \mathbf{x}_k es univariante, simplemente escribiremos $\mathbf{x}_k = x_k$.

El vector de imputación \mathbf{x}_k es instrumental para producir los valores imputados \hat{y}_k . Si la(s) variable(s) de imputación en \mathbf{x}_k son predictores importantes para la variable imputada y , podemos esperar 'imputaciones cercanas'. El error de imputación para el elemento k , $\hat{y}_k - y_k$, debería ser pequeño.

El vecino más cercano y el *hot deck* son métodos *basados en donantes*, lo que significa que el valor imputado es un valor que fue realmente observado, aunque para un elemento distinto. Esto nos asegura que el valor imputado es uno que puede ocurrir, no es un valor imposible. El *hot deck*, que veremos a continuación, es un método de imputación *aleatorio*, mientras que el vecino más cercano es *determinístico*.

La imputación por reglas estadísticas a menudo se realiza de forma automática, para un número grande de unidades, usando programas informáticos existentes. Esta imputación mecánica se realiza a menudo por *grupos de imputación*. Estos grupos tienen que ser identificados desde el principio. Un grupo de imputación se considera aquel formado por 'elementos similares'.

A menudo se imputa de acuerdo con una *jerarquía de métodos*. El método más sólido, aquél que producirá las imputaciones más 'cercanas', es el primero en aplicarse dentro de un grupo de no informantes. A continuación, si la información auxiliar necesaria para los métodos preferibles de imputación no está disponible para todos los elementos, se utiliza el segundo método más sólido al siguiente grupo, y así consecutivamente.

La imputación por una regla estadística se ve motivada por la percepción del estadístico de una relación fuerte entre la variable de estudio y y el vector de imputación \mathbf{x} . Tenemos valores observados y_k sólo para el conjunto de unidades r_i , que son las que han respondido al ítem i -ésimo para y . Por tanto, los valores imputados se calculan en el enfoque combinado para $k \in r - r_i$, usando los valores observados y_k para $k \in r_i$ y otra información.

9.8.1 Imputación por regresión

En el método de *imputación por regresión*, el valor imputado para un valor *missing* y_k es

$$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_i, \quad (9.15)$$

donde

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{k \in r_i} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in r_i} a_k \mathbf{x}_k y_k. \quad (9.16)$$

El vector de coeficientes de regresión $\hat{\boldsymbol{\beta}}_i$ es el resultado de una regresión múltiple usando los datos (y_k, \mathbf{x}_k) disponible para $k \in r_i$, y ponderado con valores debidamente especificados a_k . En el caso más sencillo, todos los a_k son iguales.

En el caso especial de una regresión lineal simple con término independiente, tenemos $\mathbf{x}_k = (1, x_k)'$, y el valor imputado es $\hat{y}_k = \bar{y}_{r_i;a} - (x_k - \bar{x}_{r_i;a})B_{r_i;a}$, donde $\bar{y}_{r_i;a} = \frac{\sum_{k \in r_i} a_k y_k}{\sum_{k \in r_i} a_k}$, $\bar{x}_{r_i;a}$ se define de forma análoga, y $B_{r_i;a} = \frac{\sum_{k \in r_i} a_k (x_k - \bar{x}_{r_i;a})(y_k - \bar{y}_{r_i;a})}{\sum_{k \in r_i} a_k (x_k - \bar{x}_{r_i;a})^2}$.

Dos casos importantes dentro de los especiales son *imputación por razón* e *imputación por media del informante*.

Imputación por razón e imputación por media del informante

Cuando $\mathbf{x}_k = x_k$ es siempre una variable de imputación positiva y unidimensional, y $a_k = \frac{1}{x_k}$, el valor imputado en (9.15) se convierte en $\hat{y}_k = x_k \hat{\beta}_i$, con $\hat{\beta}_i = \frac{\sum_{k \in r_i} y_k}{\sum_{k \in r_i} x_k}$. Esta regla de *imputación por razón* se usa a menudo cuando la misma variable se mide en dos ocasiones distintas en una encuesta repetida. Entonces y denota la variable de estudio en la encuesta actual, y x es la misma variable en una ocasión anterior. Para ilustrar esta idea, si y y x representan el 'ingreso bruto de una empresa' en dos ocasiones, entonces la 'razón actual' $\hat{\beta}_i$ mide el cambio en el nivel del ingreso de la empresa entre esas dos ocasiones.

En particular, cuando $x_k = a_k = 1 \forall k$, el valor imputado en (9.15) se convierte en $\hat{y}_k = \bar{y}_{r_i} \forall k \in s - r_i$, donde $\bar{y}_{r_i} = \frac{\sum_{k \in r_i} y_k}{m_i}$. Esto se denomina *imputación por la media del informante*. Todos los elementos que necesitan ser imputados reciben el mismo valor imputado. La distribución de los datos completos para esta variable de estudio tendrá una apariencia poco natural, con un pico en \bar{y}_{r_i} .

9.8.2 Imputación por el vecino más cercano

En el procedimiento de *imputación por el vecino más cercano*, el valor imputado para el elemento k viene dado por $\hat{y}_k = y_{\ell(k)}$, donde $\ell(k)$ es el elemento donante para el

elemento k que no ha respondido. Es decir, $\ell(k)$ proporciona su valor de y como valor imputado para el elemento k . La idea estadística que motiva este método es que dos elementos cuyos valores de x son cercanos deberían tener también valores de y cercanos.

El donante del elemento k se identifica por la minimización de una distancia como veremos a continuación. Asumiendo una variable de imputación unidimensional x , se define la distancia de un potencial donante ℓ al elemento k como $D_{\ell k} = |x_\ell - x_k|$. El donante $\ell(k)$ es el elemento que pertenece al conjunto r tal que $\min_{\ell \in r} D_{\ell k}$ se obtiene precisamente para $\ell = \ell(k)$. Es decir, las distancias $D_{\ell k}$ se calculan para todos los elementos $\ell \in r$, y el elemento donante para k será el que alcance la mínima distancia $D_{\ell k}$.

Para el elemento k , imputamos el valor y del donante, es decir, $\hat{y}_k = y_{\ell(k)}$. Como $\ell(k)$ es el más cercano a k , medido por $D_{\ell k}$, se denomina 'vecino más cercano' de k . Si el vector de imputación es multivariante, podemos minimizar una medida de distancia multivariante, por ejemplo, $D_{\ell k} = (\sum_{j=1}^J h_j (x_{j\ell} - x_{jk})^2)^{\frac{1}{2}}$, donde las cantidades h_j se especifican de forma que den un peso ajustado de los J componentes del vector de diferencia $\mathbf{x}_\ell - \mathbf{x}_k$.

9.8.3 Imputación *hot deck*

En el procedimiento de imputación *hot deck*, el valor imputado para el elemento k es $\hat{y}_k = y_{\ell(k)}$, donde $\ell(k)$ es un donante aleatoriamente elegido entre todos los elementos potencialmente donantes $\ell \in r_i$. Es un método de imputación aleatorio basado en donantes. La distribución de los valores del conjunto de datos completo resultante parecerá totalmente natural, pero todavía puede diferir considerablemente de la imagen visual obtenida a partir de la distribución (imaginada) de una muestra completa de datos de y , $\{y_k : k \in s\}$. Esto se debe a que en la imputación *hot deck*, cada donante es necesariamente un informante, y los informantes que proporcionan datos y los que no pueden ser considerablemente distintos en relación con la media, varianza y otras características.

En la imputación por regresión y por el vecino más cercano, la esperanza de tener valores imputados cercanos se basa en la hipótesis de una relación fuerte entre la variable de estudio y y el vector de imputación \mathbf{x} . La imputación por la media del que responde y la imputación *hot deck* no usan ninguna información esencialmente. Con estos métodos tan deficientes, uno corre el riesgo de imputar valores que no son 'sustitutos cercanos'. Ninguno de estos métodos es recomendable si existen alternativas mejores. En ocasiones se usan como 'métodos de último recurso', en ausencia de variables de imputación informativas. Cumplirán por lo menos uno de los objetivos de la imputación, la creación de una matriz de datos rectangular completa.

9.8.4 Grupos de imputación

La imputación a menudo se realiza dentro de *grupos de imputación* disjuntos, s_g , $g = 1, \dots, G$, cuya unión es la muestra completa s . Dentro de cada grupo de imputación la

imputación se realiza usando el mismo método. Cuando la imputación se realiza dentro del subgrupo $s_g \subset s$, usando uno de los métodos vistos anteriormente, reemplazamos s, r_i y $s - r_i$ en la fórmula apropiada por s_g, r_{ig} y $s_g - r_{ig}$, respectivamente, donde r_{ig} es la respuesta al ítem dentro del grupo g .

Podemos distinguir dos razones para usar más de un grupo de imputación en el proceso de imputación. La primera razón es que se cree que existen distintas relaciones entre los distintos subgrupos de la muestra. La relación entre y y el vector de imputación x debería de formularse teniendo esto en cuenta. Por ejemplo, si se usa la imputación por razón, la razón 'suma de y_k ' entre la 'suma de x_k ' puede diferir en los distintos subconjuntos de la muestra, sugiriendo una imputación por razón para distintos grupos. Formar un conjunto relevante de grupos requiere buen conocimiento del tema. Generalmente se suelen usar grupos por tamaño y/o actividad económica (en encuestas económicas) o por edad y sexo (en una encuesta social).

El segundo motivo es la limitada disponibilidad de variables auxiliares para imputación. La(s) variable(s) necesaria(s) para un determinado método de imputación pueden no estar disponibles para la muestra entera s . Esto puede forzar una *jerarquía de métodos de imputación*. Los métodos de imputación más estrictos se utilizan en primer lugar, en uno o más grupos y para tanta falta de respuesta como sea posible, y métodos progresivamente más débiles se utilizan en el resto de grupos. Supongamos que un vector de imputación con una fuerte relación x está disponible, pero sólo para un subconjunto de los elementos muestrales. Entonces la imputación por regresión o el vecino más cercano se puede usar con buenos resultados para ese subconjunto. A continuación, podemos tener que imputar grupos sucesivos con vectores x progresivamente más débiles. Métodos como la media del informante o *hot deck* se pueden usar como último recurso para el resto de grupos para los cuales no se disponga de ninguna o de poca información.

9.8.5 Introducción de un residuo seleccionado aleatoriamente

Tal y como se ha mencionado antes, la imputación por regresión y por razón son métodos determinísticos: dan lugar al mismo valor imputado si se repite el proceso. Sin embargo, existen ciertos motivos para que resulte de interés hacerlos estocásticos mediante la introducción de un *residuo seleccionado aleatoriamente* como ilustraremos a continuación.

Consideremos la imputación por regresión y un conjunto completo que se ha conseguido imputando $\hat{y}_k = \mathbf{x}_k^t \hat{\beta}_i$ como en (9.15) para los elementos que necesitaban imputación. Este conjunto de datos completo tiende a tener menos variabilidad que un conjunto de valores realmente observados y_k , porque el ajuste por regresión que se obtiene es, de alguna forma, el resultado de aplicar un suavizado a los datos. Añadiendo un residuo (seleccionado aleatoriamente) aliviará este problema. Alterará el aspecto del conjunto de datos completo en el sentido de que tendrá una variabilidad más natural. Como resultado, en el caso de imputación por regresión, el valor imputado del elemento k

es $\hat{y}_k = \mathbf{x}_k^t \hat{\beta}_i + e_{0k}$, donde $\hat{\beta}_i = (\sum_{k \in r_i} a_k \mathbf{x}_k \mathbf{x}_k^t)^{-1} \sum_{k \in r_i} a_k \mathbf{x}_k y_k$, como antes, y e_{0k} es un residuo seleccionado aleatoriamente de un conjunto de residuos $\{e_k : k \in r_i\}$, con $e_k = y_k - \mathbf{x}_k^t \hat{\beta}_i$.

La técnica de añadir un residuo seleccionado aleatoriamente se puede realizar (a) únicamente para estimación puntual, (b) únicamente para estimación de la varianza, o (c) para ambos. La consecuencia de (a) es que se añade varianza al estimador imputado, lo que puede verse como algo no deseable. El caso (b) representa un uso más importante de esta técnica. Si se utiliza para la estimación de la varianza, una ventaja es que el conjunto de datos completo puede entonces ser más adecuado como parte de un procedimiento de estimación de la varianza.

Comentario 2. Es necesario dedicar un comentario especial en relación con la imputación de variables cualitativas. Consideremos el caso de una variable de estudio dicotómica como la presencia o ausencia de una propiedad, como puede ser 'empleado' o 'desempleado', con valores 1 y 0, respectivamente. Para reunir los requisitos de que el valor imputado sea uno que realmente pudiese ocurrir, este debería ser 1 o 0. Una ventaja del método *hot deck* y del método del vecino más cercano es que satisface este requisito. Por contra, la imputación por regresión múltiple y sus casos especiales normalmente imputarán valores distintos de 0 o 1. Por ejemplo, en el caso más sencillo de imputar la tasas de respuesta observada dentro de los grupos, $\hat{y}_k = \frac{m_g}{n_g}$ para todos los valores *missing* en el grupo g , hemos imputado valores en el interior del intervalo unidad.

Aunque quizá 'buena en promedio', esta imputación da lugar, para cualquier elemento particular, a un 'valor imposible'. Lo mismo se verifica cuando se usa el modelo de regresión logística para obtener el valor imputado dado un elemento k como $\hat{y}_k = \exp(-\mathbf{x}_k^t \hat{\beta}_i) [1 + \exp(-\mathbf{x}_k^t \hat{\beta}_i)]^{-1}$, donde $\hat{\beta}_i$ es un vector de parámetros ajustados basado en los datos para los elementos k en el conjunto r_i . Mientras la imputación se use sólo para producir estadísticos para agregados de elementos, no hay una clara desventaja en imputar 'valores imposibles'. ■

Comentario 3. En la técnica conocida como imputación múltiple, se hacen varias imputaciones para el mismo elemento en el conjunto de datos de la encuesta. Esto contrasta con los métodos de imputación que asignan un único valor que se han discutido hasta ahora.

En la imputación múltiple, se hacen dos o más imputaciones para un valor *missing*. Esto da lugar a varios conjuntos de datos completos diferentes. Supongamos que se obtienen tres conjuntos de datos de este tipo. Los valores de y_k para los que han respondido son los mismos en los tres, pero los valores imputados (para la falta de respuesta parcial y/o la falta de respuesta total) son diferentes en los tres conjuntos. Esto asume que se ha utilizado una técnica de imputación aleatoria, por ejemplo, la imputación *hot deck*.

(Métodos determinísticos como los vecinos más cercanos y la imputación por regresión no se tienen en cuenta, porque proporcionan un valor y el mismo en las distintas repeticiones, a no ser que el método se modifique de forma adecuada).

La técnica de imputación múltiple fue propuesta por D. Rubin (véase p.ej. [Rubin 1987](#)). La imputación múltiple está diseñada tanto para la estimación puntual como para la estimación de la varianza. Una de sus principales ventajas reside en la estimación de la varianza, que se convierte en algo muy sencillo, debido a la existencia de varios conjuntos de datos completos.

En los institutos de estadística, la imputación múltiple ha tenido poco uso. Un motivo puede ser que este método demanda capacidad de almacenamiento y procesamiento de datos muy alta (aunque sólo los valores imputados difieren de un conjunto a otro). La imputación múltiple se usa en análisis secundarios de datos de encuestas. ■

Bibliografía

- Couper, M.P., R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nichols y J.M. O'Reilly, eds. (1998). *Computer assisted survey information collection*. Wiley.
- Granquist, L. (1984). "Data Editing and its Impact on the Further Processing of Statistical Data". En: *Workshop on Statistical Computing, Budapest*.
- (1995). "Improving the Traditional Editing Process". En: Wiley, págs. 385-401.
 - (1997a). "Macro-editing: a review of some methods for rationalizing the editing of survey data". En: *Statistical data editing: methods and techniques*.
 - (1997b). "On the current best methods document: edit efficiently". En: *UN/ECE Work Session on Statistical Data Editing WWP*. 30, págs. 1-8.
 - (1997c). "The new view on editing". En: *International Statistical Review* 65, págs. 381-387.
- Granquist, L. y J.G. Kovar (1997). "Editing of survey data: how much is enough?" En: Wiley, págs. 415-435.
- Leeuw, E.D. de (2005). "To mix or not to mix data collection modes in surveys". En: *Journal of Official Statistics* 21, págs. 233-255.
- Little, R.J.A. y D.B. Rubin (2002). *Statistical analysis with missing data*. 2nd. Hoboken: Wiley.
- Nordbotten, S. (1955). "Measuring the Error of Editing the Questionnaires in a Census". En: *Journal of the American Statistical Association* 50, págs. 364-369.
- (1963). "Automatic Editing of Individual Statistical Observations". En: *Conference of European Statisticians Statistical Standards and Studies No. 2*, United Nations, New York.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
- Särndal, C.-E. y Lundström S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Statistical Methodology, Federal Committee on (1990). "Data Editing in Federal Statistical Agencies". En: *Statistical Policy Working Paper 18*, U.S. Office of Management and Budget, Washington, D.C.

- UNECE (2019). *Generic Statistical Data Editing Model GSDEM*. Página visitada el día 22 de febrero de 2022. URL: <https://statswiki.unece.org/display/sde/GSDEM>.
- Waal, T. de, Pannekoek J. y Scholtus S. (2011). *Handbook of statistical data editing and imputation*. Amsterdam: Wiley.
- Wallgren, A. y B. Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.