Tema 6: Microdepuración e Imputación de Datos

En este capítulo trataremos un problema de gran importancia práctica en el procesamiento de una encuesta: la depuración e imputación de los datos. El objetivo último del capítulo será el estudio del conjunto de tareas y métodos de diseño y realización que tienen que ver con el análisis de los datos desde el punto de vista de su completitud, validez y consistencia, o dicho de otra forma, la depuración de la información obtenida. Incidiremos especialmente en el control de los posibles errores en los datos estadísticos y métodos para su tratamiento.

1.- Depuración de datos

El problema al que nos enfrentamos se presenta de formas y por causas muy variadas aunque, en términos generales, podemos decir que éste consiste en que parte de la información recogida en la encuesta falta o no es correcta o lo que es lo mismo, durante todo el proceso de recogida y tratamiento de los datos se han producido lo que hemos denominado errores ajenos al muestreo.

Recordamos que en el proceso de recogida y tratamiento de los datos estadísticos se pueden producir distintos tipos de errores. No hablamos de los errores debidos al muestreo, sino de los errores en los datos de la encuesta, es decir, los **errores ajenos al muestreo**, que a su vez clasificamos en *errores en las identificaciones*, que afectan a todo el proceso de manipulación y clasificación de la información y *errores en los datos* propiamente dichos.

Siguiendo a Granquist (1984) distinguiremos entre errores sistemáticos y errores aleatorios. Es importante resaltar que esta clasificación será importante de cara al tratamiento y análisis de los errores producidos.

Errores de tipo errático. Se producen sobre un conjunto de datos estadísticos a lo largo de su proceso de elaboración. Son producto de la falta de cuidado, luego se pueden producir en cualquier momento del proceso estadístico (uniformidad en su distribución). Al estar uniformemente repartidos, se espera que su repercusión final sobre los resultados de la estadística sea de poca importancia. No distorsionan de forma grave las distribuciones, por ello su detección y eliminación tiene una importancia menor.

Errores sistemáticos. Son aquellos que se producen por el mal entendimiento de las preguntas, conceptos o definiciones e instrucciones tanto por parte del encuestado como por parte de los individuos que intervienen en las distintas etapas del tratamiento del dato. También se incluyen aquellos producidos por respuestas intencionadamente erróneas (con vista a proteger su intimidad, por desconfianza, etc). Son más difíciles de detectar y eliminar.

Los errores sistemáticos se caracterizan, también en términos generales y con las debidas excepciones, por el hecho de que, si se repitiera la encuesta en las mismas

condiciones, se volverían a producir y muy probablemente en los mismos registros. Al contrario de los de tipo errático, al afectar a unos valores determinados de alguna de las variables, pueden distorsionar las distribuciones. Por ello, su detección y eliminación es más importante.

Llamaremos depuración de datos de una encuesta a un *conjunto de técnicas que nos permiten, a partir de la información recogida en la encuesta y, a veces, a partir de otra información adicional, corregir una parte de los errores presentes en la encuesta*.

En su forma de realización hay dos enfoques para la depuración, el enfoque tradicional, que llamaremos *microdepuración* y un enfoque más reciente conocido como *macrodepuración*. La microdepuración consiste en la detección y corrección de los errores en todos los registros de la encuesta, la macrodepuración, por su parte, consiste en hacer una selección de unidades influyentes, de valores influyentes o de ambos y una corrección de sus posibles errores. Los criterios que determinan la influencia son la representatividad del registro en la encuesta o la magnitud de los valores de los campos en los registros. La macrodepuración y la microdepuración no son procesos excluyentes en una encuesta.

1.1.- Función y Métodos

Vamos a presentar la depuración de datos desde tres ópticas distintas. La primera la dedicaremos a las funciones de la depuración, la segunda a las tareas de la depuración y por último haremos un repaso sobre los métodos de depuración. Empezamos introduciendo al lector en las funciones de la depuración.

Podemos decir que la depuración tiene tres funciones principales:

- **Facilitar el proceso de los datos**. Los procesos a los que se someten los datos en bruto hasta la obtención de resultados, como son la elaboración, almacenamiento, tabulación y análisis requieren que los datos cumplan unos requisitos mínimos. Identificadores correctos, inexistencia de duplicados, respeto del tipo y formato de las variables, etc, son requisitos a satisfacer para evitar fallos en la ejecución de los programas.

- **Mejorar la calidad de los datos.** El que la depuración mejora la calidad es un tema bastante controvertido, existiendo una fuerte polémica al respecto, como ya comentaremos más adelante.

- **Proporcionar información para medir la calidad de los datos.** La medida de la calidad de los datos se realiza en buena parte, con información producida por el proceso de depuración. La medida de la calidad de los datos es útil para:
    - Recoger información que nos permita mejorar los procesos de obtención de los datos, especialmente en futuras realizaciones.
    - Saber hasta que nivel los datos disponibles son fiables para hacer inferencias a partir de ellos. Se deben completar los datos con información adicional sobre su calidad, por ejemplo, indicando si son datos imputados u observados, etc.

El énfasis en la depuración como un proceso de control de calidad lleva a Granquist (1984) a decir que el papel de la depuración debería ser: "Principalmente, evaluar la calidad de la encuesta con información tan detallada como sea posible sobre la calidad, aún con información sobre valores de las variables…Subordinadamente, debe corregir errores que dificulten el proceso de los datos".

Una definición sobre la depuración de datos que incide sobre las tareas a realizar la facilita el grupo de trabajo de la Conferencia de Estadísticos Europeos "Data Editing Joint Group" en su glosario de términos, que definen la depuración como "Una actividad dirigida a asegurar que los datos cumplan ciertos requerimientos, es decir, satisfagan condiciones de corrección establecidas". Habitualmente se compone de tres fases:

1. La definición de un sistema de requerimientos consistente.
2. Su verificación sobre los datos
3. La eliminación o sustitución de los datos que están en contradicción con los requerimientos"

Un punto a destacar en esta definición es que no se plantea como un objetivo de la depuración que los datos manejados sean "verdaderos" o "exactos", lo cual sería un objetivo imposible. Se limita a pedir una, mucho más modesta, satisfacción de ciertos requerimientos.

En cuanto a los métodos de la depuración, en su forma de realización hay dos enfoques:

- El enfoque tradicional, que llamaremos **microdepuración,** consistente en la detección/corrección de los errores en todos los registros de la encuesta.

- Un enfoque más reciente, conocido como **macrodepuración,** consistente en hacer una selección de las unidades influyentes, de valores influyentes o de ambos y una corrección de sus posibles errores. Los criterios que determinan la influencia son la representatividad del registro en la encuesta (es decir su peso, o que corresponde a una unidad muestral muy importante), o la magnitud de los valores de los campos de los registros. Como veremos, la macrodepuración y la microdepuración no son procesos excluyentes en la depuración de una encuesta.

1.2.- Coste y Críticas

Las tareas descritas en el apartado anterior tienen un coste económico, y un coste en tiempo que ocasiona retrasos en la presentación de resultados. En la evaluación del coste económico se deberían incluir las siguientes partidas:

- El diseño del método de depuración.
- La elaboración de los procedimientos de revisión de cuestionarios previa a la grabación.
- La formación de los depuradores.
- La elaboración de los programas para los procesos de depuración.
- El coste del personal dedicado a la corrección de los errores detectados.
- El tiempo de ordenador.

En la mayoría de los casos no se dispone de una contabilidad detallada que permita calcular que porcentaje de los gastos de la producción de las estadísticas corresponde a la depuración. El Subcommittee on Data Editing in Federal Statistical Agencies del Federal Comité on Statistical Methodology de EEUU considera una estimación válida de este coste de al menos un 20% del gasto total para la depuración en estadísticas sociodemográficas y un 40% para estadísticas económicas.

En cuanto a las críticas, el proceso de depuración de datos no está exento de ellas. Se cuestionan los gastos que los procesos de detección y corrección conllevan y los procesos de imputación. Es indudable que en los casos en que la detección de un error da lugar a la obtención del verdadero valor, bien al descubrirse un error en la grabación o la codificación, estando el dato verdadero en el cuestionario, bien porque se accede de nuevo al informante, la depuración mejora la calidad del dato.

La controversia está en que si la mejora justifica los gastos en montar los procedimientos y mantener los equipos de depuradores. Actualmente se pone énfasis en la investigación de métodos que eviten tareas manuales, si no totalmente, si reduciéndolas a la detección y eliminación de errores realmente importantes. La creencia actualmente más extendida parece ser la de que es necesaria alguna depuración, aunque sin precisar su extensión y forma. A continuación se citan algunos de los argumentos más empleados a favor y en contra de la depuración de datos.

- **Argumentos a favor de la depuración**:
  - Facilita las fases posteriores del tratamiento de la encuesta. Por ejemplo, si en estas fases hubiera valores perdidos o fuera de rango, los estadísticos obtenidos tendrían menos fiabilidad.
  - Los lectores de resultados de encuestas, al encontrarse el "no sabe" o "no contesta" o, más en general, la información que falta, la eliminan mentalmente y piensan en términos del resto de categorías. Esto equivale a asignar esta información que falta, proporcionalmente, al resto de categorías y constituye un tratamiento de la información que falta bastante elemental. Parece lógico, entonces, que si el usuario va a realizar este tratamiento elemental, el analista pueda realizar antes otro más perfecto.
  - Los datos depurados inspiran confianza al lector. En efecto, algunos lectores, desconocedores de la complejidad del tratamiento estadístico de datos, desconfiarán si ven, por ejemplo, una tabla donde aparezca un titulado superior menor de 15 años.
- **Argumentos en contra de la depuración:**
  - A veces, cuando existe una inconsistencia en la que están implicadas varias variables, es difícil decir cuál es la errónea y se producen errores adicionales.
  - La depuración puede incitar a que se descuide la recogida de la información, con la esperanza de que aquella corrija a ésta, cosa que no sucederá.
  - La depuración produce datos "bonitos", que dan una falsa sensación de corrección.
  - Se puede perder información válida, particularmente en las zonas fronteras de las reglas de detección.

- La falta de respuesta puede no ser aleatoria; si se procede a realizar imputaciones de registros enteros para subsanarla, estamos introduciendo sesgos en las estimaciones.
- En el caso de imputar casos de falta de respuesta total se están "fabricando" datos.
- Un procedimiento de imputación basado en supuestos poco realistas o con una metodología pobre, puede hacer que la calidad de los datos empeore sustancialmente.

En resumen, tanto los argumentos a favor como en contra tienen partes de verdad, por lo cual, deberemos en todo caso cuidar la recogida de la información e intentar dar buenos criterios de depuración, que no perjudiquen más que beneficien a los datos.

Finalmente, hemos de insistir en que los dos motivos principales para realizar una depuración. El primero de ellos es mejorar la calidad de los datos. Como acabamos de ver, este es un objetivo cuyo cumplimiento produce controversias. El segundo objetivo que destacaremos el conocer la calidad de los datos. Este objetivo es menos conflictivo que el anterior, pues parece claro que la depuración sí permite cumplirlo y que ese cumplimiento es deseable en todos los casos.

2.- Microdepuración de Datos

La microdepuración se compone de las siguientes actividades básicas:
- Detección de los registros sospechosos.
- Identificación de las variables a modificar.
- Asignación de nuevos valores a las variables identificadas como erróneas.

Pudiendo realizarse esta asignación de dos maneras:
- Mediante corrección manual.
- Mediante imputación.

Las fases de detección, identificación y corrección manual pueden diseñarse como un proceso cíclico. La imputación automática debe ser una etapa final y única del proceso de depuración de datos, en el sentido estricto de su definición.

En este epígrafe analizamos la detección de registros erróneos, presentando la herramienta habitualmente utilizada: los edits. Dedicamos también un apartado a la selección de las variables a ser modificadas y otro a la corrección manual. Por último estudiaremos la imputación automática desde el punto de vista teórico.

2.1.- Definición de Edits. Análisis

La detección de errores requiere que previamente se definan las situaciones erróneas o sospechosas; esta tarea corresponde a los estadísticos expertos en el tema objeto de la investigación. La definición de posibles situaciones erróneas se realiza habitualmente por medio de los que se denominan edits. Los edits especifican restricciones a los valores individuales de las variables (edits de validación), o de conjuntos de variables (edits de consistencia). Estas restricciones se pueden especificar de manera positiva como reglas de aceptación (edits de aceptación), o de manera negativa como reglas de

rechazo (edits de conflicto). Aplicando operadores sencillos es inmediato pasar de los edits de conflicto a los de aceptación y viceversa.

La detección de errores se realiza enfrentando todos los registros a depurar con el conjunto de los edits especificados. Un registro se considera erróneo si cumple la condición especificada por un edit de conflicto, o si incumple la condición especificada por un edit de aceptación.

Mediante los edits se especifican:

- **Situaciones imposibles**: Hay valores de variables, o combinaciones de valores, que son imposibles en la realidad. Por ejemplo, una madre no puede ser más joven que su hijo, una parcela no puede ser más grande que el término municipal en el que está enclavada, etc. Estas situaciones deben ser detectadas y eliminadas.

- **Situaciones improbables**: Son situaciones de difícil aparición en la realidad, pero que no se pueden rechazar de entrada, por ejemplo, madres de 14 años o con más de 15 hijos. Es de interés poder detectar estas situaciones, aunque el problema esta en definirlas de manera precisa, y de forma que no nos veamos inundados por una avalancha de presuntas situaciones raras, que luego no lo son tanto.

- **Restricciones contables**: En datos económicos existen restricciones contables a las que los datos deben ajustarse.

- **Outliers estadísticos**: En variables numéricas puede que aparezcan valores que no se ajustan a la distribución de los otros datos, bien porque corresponden a unidades influyentes o bien porque son valores erróneos. Los edits utilizados para la detección de outliers se denominan edits estadísticos.

- **Control del flujo de respuesta en el cuestionario**: La mayor parte de los cuestionarios tienen reglas que dirigen el flujo de respuesta a través del mismo. Los sistemas CAPI y CATI aseguran el flujo correcto de respuestas. En caso de no utilizar uno de estos métodos, se deben utilizar edits de flujo.

Según el tipo de variables a las que se apliquen, distinguimos entre edits de tipo cualitativo y edits de tipo cuantitativo. Los edits de tipo cuantitativo se aplican a variables numéricas y generalmente se expresan como restricciones de rango (edits de validación) o como igualdad o desigualdad entre funciones de las variables (edits de consistencia). Los actuales sistemas generales de depuración de datos trabajan exclusivamente con expresiones lineales.

Los edits de tipo cualitativo se aplican a variables categóricas. Los edits de validación se expresan con la lista de códigos válidos. Los edits de consistencia, generalmente de rechazo, se expresan como combinaciones de códigos de dos o más variables que son inaceptables o sospechosas.

Otro tipo de edits son los edits condicionales que añaden al edit (cuantitativo o cualitativo) una condición que determina el subconjunto de registros a los que se aplica tal edit. Son de la forma:

IF condición THEN aplicar edit

Cuando los edits condicionales incluyen una condición cualitativa y un edit cuantitativo se trabaja con los dos tipos de variables, numéricas y categóricas, lo cual dificulta su tratamiento en los sistemas generales.

En lo que se refiere al análisis, podemos considerar el conjunto de todas las posibles respuestas a las preguntas del cuestionario, como un espacio de dimensión igual al número de variables (o preguntas), por lo que en el análisis de los edits estudiaremos las restricciones a este espacio. El conjunto de todos los edits define la región de aceptación.

Puede ocurrir que un edit no aporte nada a la definición de la región de aceptación, pues es redundante respecto a los demás edits. Los edits redundantes deben ser eliminados para facilitar la tarea de detección de errores. Un problema más grave es que el conjunto de edits defina una región de aceptación demasiado restringida, siendo los casos más extremos cuando la región de aceptación es vacía, o se reduce, en algún sentido, la dimensión del espacio. En estos casos decimos que el conjunto de edits es inconsistente, siendo necesario modificar o eliminar uno o más edits.

Se debe realizar un análisis de los edits antes de enfrentarlos a los registros a depurar, pues si existe una inconsistencia no detectada puede tener un impacto importante al rechazar registros válidos, y si se detecta durante el procesamiento de los datos, es necesario rehacer los procesos produciéndose retrasos.

2.2.- Campos a Modificar. Asignación de Valores

Para cada registro detectado como erróneo (que falla uno o más edits), es necesario identificar los campos a modificar para corregir el error. Los campos con valor fuera de rango o con códigos inválidos no suponen un problema de identificación. El problema surge en el caso de inconsistencias entre valores individualmente válidos. Por ejemplo, el registro: EDAD: 8 años y ESTADO CIVIL: viudo.

EL proceso debe seleccionar al menos un campo que figure explícitamente en cada edit fallado, existen varias alternativas para ello. Una primera alternativa es seleccionar todos los campos que intervienen en todos los edits fallados. Pero ello supondría una gran pérdida de información. Otros criterios habitualmente considerados son:

- Escoger el menor número posible de campos que cubran todos los edits fallados (principio de cambio mínimo, propuesto por **Fellegi y Holt**).

- Escoger campos que cubran todas las reglas de conflicto, de forma que se minimice la suma de pesos asociados a los campos del registro. Estos pesos indican el grado de confianza que el experto tiene en el valor de las variables del registro.

- Escoger las variables que faciliten una imputación más fácil.

El proceso de identificación, también llamado de localización de errores, se puede realizar por un operador humano o automáticamente.

El paso siguiente será asignar a cada campo identificado en la etapa anterior como erróneo un valor válido y consistente con el resto de los valores del registro. El proceso de asignación se puede realizar manualmente, y lo denominaremos "corrección manual" o automáticamente, y lo denominaremos "imputación". En el caso de corrección manual, el valor a asignar puede obtenerse de:

- El propio cuestionario si el error se produjo en una fase posterior a la cumplimentación, o si el cuestionario tiene notas al margen o información no grabada que nos permita inferir el verdadero valor, o si aprecia un error obvio, por ejemplo en las unidades de medida.

- Contactar nuevamente con la fuente que suministro la información. En todos los demás casos debe ser un experto el que facilite el valor del campo o se debe dejar la asignación a la imputación automática.

En el caso de imputación automática el valor a asignar se obtiene mediante un proceso de estimación determinística o aleatoria, que denominamos respectivamente imputación determinística e imputación probabilística. Los procedimientos de imputación automática serán descritos en el epígrafe siguiente.

La fase de detección de errores con corrección manual es muy costosa en tiempo, pudiendo, si no se realiza adecuadamente, ser fuente de sesgos en los datos finales. Algunos aspectos a considerar para evitar estos problemas son:

- La detección de errores con corrección manual está asociada a la idea de un proceso cíclico: detección/corrección, en muchos casos interminable. Los ordenadores han aportado la posibilidad de aplicar con uniformidad y sin errores un mayor número de edits, pero los operadores humanos se pueden ver desbordados por el gran número de tipos de error a los que se pueden tener que enfrentar. Muchos de esos errores son de muy poca importancia.

- Evitar que los edits definan regiones de aceptación demasiado reducidas, que requieren un exceso de revisiones de situaciones que, o bien son correctas, o bien tienen errores poco importantes.

- La corrección manual, si no es bien informada, puede ser tan arbitraria como los métodos automáticos, sin tener ni la uniformidad ni la precisión de estos.

El proceso de microdepuración de los datos en la base primaria culmina con la imputación bien manual, bien automática de valores a campos erróneos. Es fundamental en este sentido el marcado de dichos campos dentro de los registros correspondientes mediante ciertos códigos de falta de respuesta o de borrado a fin de facilitar la tarea de búsqueda.

3. Introducción a la Imputación de Datos

En una investigación estadística, tanto parcial como exhaustiva, es frecuente que individuos encuestados no respondan a una o más preguntas del cuestionario. Cuando esto ocurre se dice que se tienen datos ausentes o missing y estamos bajo un problema de no-respuesta.

La no-respuesta puede introducir sesgo en la estimación e incrementar la varianza muestral debido a la reducción del tamaño muestral. La imputación de datos es la etapa final del proceso de depuración de datos, tras el proceso de edición, en el cual los valores missing o que han fallado alguna regla de edición del conjunto de datos son reemplazados por valores aceptables conocidos. La razón principal por la cual se realiza la imputación es obtener un conjunto de datos completo y consistente al cual se le pueda aplicar las técnicas de estadística clásicas.

Para la aplicación de la imputación de datos se recibe de la etapa anterior un fichero de datos con ciertos campos marcados por "falta de respuesta" ó "borrados" en la fase de edición por no cumplir alguna regla de edición propuesta. Tras la imputación de todas las variables del estudio se obtiene un fichero completo.

Encontrar un buen método de imputación es una tarea importante ya que errores cometidos en las imputaciones de datos individuales pueden aparecer aumentados al realizar estadísticas agregadas. Por todo esto parece razonable estudiar métodos de imputación que conserven características de la variable como pueden ser: conservación de la distribución real de la variable, relación con el resto de variables en estudio, etc. Los métodos de imputación para datos faltantes varían según el tipo del conjunto de datos, extensión, tipo de no-respuesta, etc.

Denominaremos a los campos marcados para imputar **como campos a imputar** y al registro con campos a imputar como **registro a imputar.** Recordemos que el principal objetivo de la imputación es producir un fichero de datos completo y consistente después del proceso de estimación de los campos a imputar (gráficamente, si un registro de q variables es un punto en el espacio R^q dimensional, el proceso de imputación consiste en "devolver" a la región de aceptación de la encuesta los puntos que están fuera de ella).

Otro objetivo importante que nosotros damos a esta etapa están en relación con la tercera función que asignábamos al proceso de depuración de los datos y que era la de medir la calidad de los datos de la encuesta. En este sentido, la fase de imputación debe producir un fichero que contenga, por registro de la encuesta, un conjunto de **códigos de estado** que informen para cada campo si el dato tiene valor observado o imputado y, si el dato es imputado, del procedimiento de imputación utilizado. También debería informar, si es factible y se utilizan métodos de imputación con registros donantes, del número de veces que un registro o campo fue utilizado para "donar" sus valores a un registro a imputar. Este fichero será un fichero para el control de calidad de los datos y debería explotarse y distribuirse al tiempo que se explota y distribuye el fichero de datos.

La imputación de los datos no es un proceso unánimemente aceptado por los investigadores. Se habla de la imputación como un proceso de "fabricación de datos" (véanse, por ejemplo, las críticas de Banister (1980) al proceso de imputación). Hay

razones, sin embargo, que apoyan el uso generalizado de procedimientos de imputación como son:

1. Reducir el sesgo de las estimaciones.
2. Facilitar los procesos posteriores de análisis de los datos.
3. Facilitar la consistencia de los resultados entre distintos tipos de análisis.
4. Enriquecer el proceso de estimación con fuentes auxiliares de información de la que normalmente se dispone en las Oficinas de Estadística; ejemplo de fuentes auxiliares son los datos administrativos, encuestas relacionadas con la que se procesa, etc.

En su crítica, los que cuestionan la imputación han incentivado el estudio de métodos de estimación eficientes, así como el que se controle la calidad de los datos y del proceso por medio de códigos de estado o instrumentos similares. Existen, por tanto, ciertas reticencias en el ámbito de los técnicos en estadística en torno a la aplicación de éstos métodos por lo que tienen de manipulación de la información originaria, pero hay otras muchas razones que apoyan un uso razonado de las mismas.

3.1.- Supuestos de los Procedimientos de Imputación

Al estimar los valores de los campos a imputar en base al valor de 1os restantes campos del fichero, el supuesto implícito que se hace es que el comportamiento de las unidades cuyos registros están incompletos es el mismo que el comportamiento de las unidades cuyos registros tienen dato en todos sus campos. Analizamos más detenidamente este supuesto.

El fichero a imputar tiene registros con campos sin respuesta y campos borrados por tener un valor erróneo o inconsistente. Los errores a su vez, fueron clasificados entre errores aleatorios o sistemáticos. Para los campos con errores sistemáticos es bastante cuestionable aplicarles el supuesto implícito anterior. Por ello, el experto debe detectar en depuración los posibles errores sistemáticos de la encuesta y aplicar a los mismos, si procede, lo que denominamos una **imputación determinística.**

Una imputación determinística toma generalmente el formato:

IF (condición) **THEN** (acción).

Se supone que los campos a imputar restantes son campos con **errores o falta de respuesta aleatoria**. Para estimar los valores de estos campos aplicamos los procedimientos de **imputación probabilística** que vamos a estudiar en este capítulo. Estos procedimientos exigen hacer **explícitamente** los supuestos siguientes sobre los campos con dato y campos a imputar.

1. **Supuesto sobre la falta de errores de contenido.** Esto es, los campos del registro que no son objeto de imputación (campos con dato) tienen el valor que se quiere observar; es decir, no hay errores de contenido. Formalmente, el supuesto dice:

   Sea $Y_i = (y_{i1}, y_{i2}, \ldots\ldots y_{iq})$ el vector de valores VERDADEROS a observar.
   $\{i.\ i = 1,2,\ldots,n\}$ índice de observación.

Sea $X_i = \{X_{id}, X_{im}\} = (X_{i1}, X_{i2}, \ldots\ldots\ X_{iq})$ el vector de valores OBSERVADOS donde $_d$ y $_m$ señalan los campos con dato y a imputar respectivamente. Entonces:

$$X_{id} = Y_{id}.$$

2. **Supuestos sobre los campos a imputar.** Estos supuestos definen los posibles modelos para explicar el comportamiento de las unidades cuyos registros tienen campos a imputar; denominamos a estos modelos los "mecanismos de generación de los campos a imputar". Hay tres tipos de mecanismos (los nombres de los supuestos los tomamos de Rubin & Little (1987), que definen los supuestos MAR "Missing At Random" y MCAR "Missing Completely At Random").

   I. **MCAR, Missing Completely At Random (faltante completamente aleatorio)**. Se da este tipo cuando la probabilidad de que el valor de una variable Xj , sea observado para un individuo i no depende ni del valor de esa variable, x_{ij}, ni del valor de las demás variables consideradas, $x_{ik}\ \ j \neq k$. Es decir, la ausencia de información no está originada por ninguna variable presente en la matriz de datos.

   II. **MAR, Missing At Random (faltante aleatorio):** Se da este tipo si la probabilidad de que el valor de una variable Xj sea observado para un individuo i no depende del valor de esa variable x_{ij} ,pero quizá sí del que toma alguna otra variable observada $x_{ik}\ \ j \neq k$. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos.

   III. **NMAR, No missing at Random.** Se produce este tipo de mecanismo en el caso en el cual la probabilidad de que un valor x_{ij} sea observado depende del propio valor x_{ij} , siendo este valor desconocido. En el ejemplo anterior, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores.

Generalmente, los supuestos anteriores de MAR y MCAR para el conjunto de la encuesta son difícilmente sostenibles, en cambio para el caso de realizar la imputación basada en estratos o grupos, dentro de éstos sí es mas acertado suponer los modelos MAR y MCAR. Esta es una de las causas para que las imputaciones tiendan a hacerse dividiendo la población en subgrupos disjuntos.

Visualicemos las definiciones anteriores con un ejemplo: consideremos una encuesta que recoge, por ejemplo, información sobre los INGRESOS y la EDAD de los miembros de los hogares. Si analizamos la variable INGRESOS individualmente, y observamos que un registro tiene el campo de INGRESOS en blanco, podemos pensar que la falta del dato está o no está relacionado con el verdadero valor de los ingresos del miembro del hogar:

- Si suponemos que la falta del dato no está relacionado con el verdadero valor estamos haciendo un supuesto MAR: **La falta de respuesta** en INGRESOS es aleatoria.
- Si analizamos las variables INGRESOS y EDAD conjuntamente, y suponemos que la falta de respuesta en el campo INGRESOS es independiente del verdadero valor de los ingresos del miembro del hogar y de la edad, estamos haciendo un supuesto MCAR: **La falta de respuesta es completamente aleatoria**; sin embargo, si suponemos que la falta de respuesta en el campo INGRESOS es independiente de los ingresos verdaderos del miembro del hogar pero que pueden depender de su edad, entonces hacemos un supuesto MAR: **La falta de respuesta es aleatoria.**
- Si suponemos que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores, hacemos un supuesto NMAR.

En términos generales:

Supuesto MCAR: X_{mi} es independiente de Y_{mi} y de X_{di}.
Supuesto MAR: X_{mi} es independiente de Y_{mi}

Para el conjunto de la encuesta, los supuestos anteriores pueden ser difícilmente sostenibles. Sin embargo, si se divide la encuesta en subgrupos o estratos, y se aplican a cada estrato, los supuestos MAR y MCAR son más defendibles. Esta es la causa de que los procedimientos de imputación se apliquen, generalmente, en estratos. Evidentemente, las suposiciones que se establezcan para una determinada subpoblación (estrato o conglomerado) pueden no ser válidas para otras, en la medida en que las relaciones de dependencia no suelen ser uniformes.

3.2.- Métodos de Imputación

La solución al problema del sesgo de las estimaciones consiste en imputar los datos faltantes, sustituyéndolos por valores estimados mediante algún método de imputación. Durante las décadas anteriores se empleaban procedimientos de imputación basados en la experiencia, la intuición y la oportunidad. Se suponía uniforme la probabilidad de que las unidades respondiesen y se ignoraba frecuentemente el sesgo causado por la no-respuesta.

Actualmente se emplean infinidad de métodos de imputación y se generan nuevos métodos empleando distintas técnicas estadísticas. Gran parte de los métodos de imputación se pueden expresar mediante la siguiente fórmula:

$$y_{vi} = f(y_{nm}) + \varepsilon$$

donde *vi* y representa el valor imputado, y_{nm} representa las observaciones con valores válidos (no missing), mientras que el ε se refiere al residuo aleatorio. En el caso de métodos determinísticos se asigna ε =0 y es variable en el caso de métodos estocásticos. Los primeros proporcionan mejores resultados si se tiene en cuenta los estimadores

puntuales como la media, mediana, etc., sin embargo provocan distorsiones en la distribución de la variable.

A continuación se comentan las características de los principales métodos de imputación (salvo el de la **deductiva o lógica** vista anteriormente) junto con las ventajas y desventajas de cada uno de ellos, siguiendo la clasificación propuesta por Laaksonen (2000).

3.2.1 Imputación mediante registro donante

Son procedimientos que asignan a los campos a imputar de un registro el valor que en tales campos tiene otro registro de la encuesta. A los registros completos se les denomina registros donantes y los registros con campos a imputar se denominan registros receptores o candidatos. A los campos que se utilizan para establecer la relación entre registro donante y candidatos se les denominan campos de control.

Dichos campos pueden ser tanto cualitativos como cuantitativos o de ambos tipos. En el caso de tratar variables exclusivamente cualitativas, el cruce de las distintas variables para dividir la población en subgrupos disjuntos se denominan estratos y la relación entre los registros candidatos y los donantes se establecen por igualdad de los códigos del estrato. Entre las ventajas de estos métodos se pueden destacar:

- Se imputa un valor posible y realizado.
- Es sencillo de implementar. Mientras que el principal problema se debe a que puede no haber respondientes con todo el rango de valores necesario en la variable a imputar.

Existen gran número de métodos entre los que se destacan los siguientes:

Procedimiento Cold-Deck. Se define un registro donante por estrato como "registro tipo" en base a fuentes de información externas: datos históricos, distribuciones de frecuencias, etc… El método asigna a los campos a imputar de todos los registros candidatos los valores del registro donante correspondiente al mismo estrato. A partir de este método se originó el procedimiento hot-deck. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible.

Procedimientos Hot-deck. Este método es un procedimiento de duplicación. Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo. Todas las unidades muestrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Se está suponiendo que dentro de cada grupo la no-respuesta sigue la misma distribución que los que responden. Este supuesto incorpora una fuerte restricción al modelo, si esta hipótesis no es cierta se reducirá sólo en parte el sesgo debido a la no-respuesta. El método Hot-deck tienen ciertas características interesantes a destacar:

- Los procedimientos conducen a una post-estratificación sencilla.
- No presentan problemas a la hora de encajar conjuntos de datos.
- No se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que falten.
- Preservan la distribución de la variable.

Sin embargo estos métodos tienen algunas desventajas:

- Distorsionan la relación con el resto de las variables.
- Carece de un mecanismo de probabilidad.
- Requieren tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza.
- Las clases han de ser definidas en base a un número reducido de variables, con la finalidad de asegurar que habrá suficientes observaciones completas en todas las clases.
- La posibilidad de usar varias veces a una misma unidad que ha respondido.

Algunas variaciones del método son:

1. **Procedimiento Hot-deck secuencial.** El registro donante es el registro sin valor missing, perteneciente al mismo estrato e inmediatamente anterior al registro candidato. Para aplicar esta imputación previamente se debe clasificar el fichero de tal forma que produzca una autocorrelación positiva entre los campos sujetos a imputación, de esta forma se asegura una mayor similitud entre registro donante y candidato. Las desventajas de este método son considerables:
   - Hay que facilitar valores iniciales para el caso de tener valores missing en el primer registro.
   - Ante una racha de registros a imputar, se emplea el mismo registro donante.
   - Es difícil de estudiar la precisión de las estimaciones.

2. **Procedimiento Hot-deck con donante aleatorio.** Consiste en elegir aleatoriamente a uno o varios registros donantes para cada registro candidato. Hay diferentes modificaciones de este método. El caso más simple es elegir aleatoriamente un registro donante e imputar el registro candidato con dicha información. Se puede elegir una muestra de registros donantes mediante distintos tipos de muestreo e imputar al valor medio obtenido con todos ellos. Este último tipo tiene un elemento de variabilidad añadida debido al diseño de elección de la muestra que incorporan.

3. **Procedimiento Hot-deck modificado.** Consiste en clasificar y encajar los donantes potenciales y receptores utilizando un considerable número de variables. El encaje se hace sobre bases jerárquicas del siguiente modo: si no se encuentra un donante para encajar con un receptor en todas las variables de control, se eliminan algunas variables consideradas como menos importantes y de esta forma conseguir el encaje a un nivel superior.

Procedimiento DONOR. En este método se emplea una función distancia definida entre las variables para que se mida el grado de proximidad entre cada posible registro donante y el registro candidato. En este caso se imputa en bloque los valores del registro donante en los campos sin respuesta del candidato. Es necesaria una modificación previa de los datos para anular los efectos de escala en la función distancia.

3.2.2.- Imputación mediante modelos donantes

Son procedimientos que asignan a los campos a imputar de un registro valores generados a partir de modelos ajustados a los valores observados de los registros

respondientes. Existen diversos métodos de imputación, los principales se comentan a continuación:

Procedimientos de regresión. Se incluyen en dicho grupo aquellos procedimientos de imputación que asignan a los campos a imputar valores en función del modelo:

$$y_{vi} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

donde y_{vi} es la variable dependiente a imputar y las variables $\{x_j | j \equiv 1 \ldots n\}$ son las regresoras que pueden ser tanto cualitativas como cuantitativas, generalmente variables altamente correladas con la dependiente. Las variables cualitativas se incluyen en el modelo mediante variables ficticias o dummy. En este tipo de modelos se supone aleatoriedad MAR, donde ε es el término aleatorio. A partir de este modelo se pueden generar distintos métodos de imputación dependiendo de:

1. Subconjunto de registros a los que se aplique el modelo.
2. Tipo de regresores
3. Los supuestos sobre la distribución y los parámetros del término aleatorio ε .

- **Imputación de la media.** El modelo basado en la imputación de la media es el modelo más sencillo de los pertenecientes a los procedimientos de regresión. Sigue el siguiente modelo:

  $$y_{vi} = \alpha + \varepsilon$$

  Este método de imputación es muy sencillo y consiste en la asignación del valor medio de la variable a todos los valores missing de la población o el estrato según se haga la imputación global o a partir de subgrupos construidos a partir de las categorías de otras variables que intervienen en el estudio. En la versión estocástica se incluye un residuo aleatorio. Este método tiene como desventajas que modifica la distribución de la variable reduciendo la varianza de la variable, como consecuencia en el caso de realizar análisis bivariantes reduce la covarianza entre las variables. Es decir, este método no conserva la relación entre las variables ni la distribución de frecuencias original. Además en este modelo se supone estar bajo un procedimiento MCAR.

- **Modelos de regresión aleatoria.** Resuelven el problema de la distorsión de la distribución tras la imputación. Se propone añadir una perturbación aleatoria a las estimaciones del modelo de regresión:

  $$\hat{x}_{im} = a + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + \hat{e}_i$$

  donde las perturbaciones $\hat{e}$ se calculan mediante alguno de los siguientes métodos:

  1. Se obtiene una muestra aleatoria de tamaño s de los r residuos observados $\hat{e}_i = \hat{x}_{im} - x_{im}$ y se suman a los valores x_{im} estimados.

  2. Se obtienen aleatoriamente s valores de una distribución con media cero y varianza $\hat{\sigma}^2$, donde $\hat{\sigma}^2$ es la varianza residual correspondiente a los valores observados de x_{im} .

Los modelos más generales de regresión tienen ciertas mejoras con respecto a la imputación de la media que se comentan a continuación:

1. Asume el supuesto menos estricto de aleatoriedad, modelo MAR.
2. Infraestima el valor de la varianza y covarianza en menor medida que en el caso de imputación a la media..
3. Modifica en menor medida la distribución de las variables.

- **Método de imputación mediante regresión logística.** Método de imputación similar al de regresión aplicable a variables binarias. Se realiza con los registros respondientes una regresión logística y en base a esta regresión se imputan los registros con no-respuesta. De la misma forma que en otros métodos está la versión determinística y la aleatoria que incluye una perturbación aleatoria.

  Recientemente se está aplicando el método de regresión logística basada en técnicas de registros donantes, con la idea de imputar los registros sin respuesta mediante los registros respondientes, implementado en el programa de imputación SOLAS (1999).

- **Método Regression-based nearest neighbour hot decking (RBNNHD).** Método propuesto por Laksonen (2000) que combina la imputación mediante métodos de regresión con los métodos de ficheros donantes. Consiste en construir una regresión lineal multivariante con los registros con respuesta. Clasificar los registros con no respuesta añadiéndoles un termino error y posteriormente ordenarlos según el valor imputado. Tras esto, se aplica la regla del vecino más próximo a los registros con valor imputado y se modifica por el asignado mediante este método donante.

3.2.3.- Métodos basados en factorizar la verosimilitud

Algoritmo EM (Expectation Maximization). Little & Rubin (1987). Método basado en factorizar la función de verosimilitud que permite obtener estimaciones máximo verosímiles (MV) de los parámetros cuando hay datos no completos con unas estructuras determinadas. Es válido para cualquier estructura de datos no completos. El algoritmo EM permite resolver de forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos en cada iteración:

1. Paso E (Valor esperado): Calcula el valor esperado de los datos completos basándose en la función de verosimilitud.

2. Paso M (Maximización): Se asigna a los datos missing el valor esperado obtenido en el paso anterior (E) y entonces se calcula la función de máxima verosimilitud como si no existiesen valores missing. Ambos pasos se realizan de forma iterativa hasta obtener convergencia.

Explícitamente consiste en resolver las ecuaciones de verosimilitud:

$$S(\Theta \mid X_d) = \frac{\delta}{\delta\Theta} (\Theta \mid X_d) = 0$$

La solución de tales ecuaciones no siempre es posible obtenerse de un modo directo (en general nunca lo es). Por ello se usan métodos iterativos como el conocido de Newton-

Raphson. El ALGORITMO EX, propuesto en Dempster et al (1977), es un método iterativo alternativo que supone una aproximación "estadística" en vez de "numérica" al problema y está especialmente adecuado al problema de los datos incompletos.

Cada iteración, t, del ALGORITMO consiste, como hemos comentado anteriormente, en un Paso-E y en un Paso-M. El algoritmo está basado en la idea intuitiva de estimar los blancos y reestimar. Las ventajas del ALGORITMO EM son su relativa facilidad de implementación y su interpretación estadística. Se puede probar que cada iteración del ALGORIT'MO aumenta la verosimilitud, hasta su convergencia. La convergencia, sin embargo, puede ser muy lenta, dependiendo de la fracción de campos a imputar.

El ALGORITMO EM esta implementado de modo general en BMDP (y por tanto en SPSS módulo de valores faltantes) MBSPEER de R. Little (1988) en una rutina FORTRAN y en el sistema CIDAC.

Algoritmo de aumento de datos. Es un procedimiento iterativo que permite obtener valores simulados de los datos ausentes y de los parámetros desconocidos, para algunas clases de modelos multivariantes. De la misma forma que el algoritmo EM, trata de solucionar un problema difícil con datos incompletos resolviendo repetidas veces problemas accesibles con datos completos. Consiste en un proceso iterativo que tiene dos fases:

1. Paso I de imputación de los datos ausentes. Simula valores para los datos ausentes mediante la distribución obtenida en la fase anterior y los valores observados.

2. Paso P o posteriori, que consiste en simular nuevos valores de los parámetros a partir de la distribución a posteriori condicionada a los datos completados en la fase anterior.

Muestreo de Gibbs y otros métodos. El muestreo de Gibbs es otro procedimiento para estimar los parámetros del modelo e imputar los datos ausentes. Se emplea el muestreo de Gibbs cuando se modela el problema de falta de datos mediante una metodología bayesiana.

Por su parte, recientemente se han propuesto, y se siguen estudiando, diversas técnicas estadísticas aplicadas en la fase de imputación de datos, como pueden ser:

- Imputación de datos basadas en distintas técnicas de redes neuronales. Que está siendo estudiado actualmente en el proyecto EUREDIT.
- Imputación basada en árboles de clasificación y regresión. Propuesta en el proyecto europeo AUTIMP como una técnica adecuada de imputación.
- Imputación basada en la lógica difusa. Técnica propuesta en el proyecto europeo EUREDIT y se está desarrollando en la actualidad.

3.3.- Estrategias de Imputación

Antes de realizar la imputación surge el problema de qué criterios se deben tener en cuenta para seleccionar el modelo de imputación a aplicar. Esta respuesta no es sencilla y hay que tener en cuenta los siguientes cinco aspectos que se detallan a continuación:

1. **La importancia de la variable a imputar.** Si la variable es de elevada importancia, es natural que se elija más cuidadosamente la técnica de imputación a aplicar.

2. **Tipo de la variable a imputar.** Hay que considerar en este contexto el tipo de la variable, es decir, si es continua ó categórica tanto nominal como ordinal. Teniendo en cuenta para el primer grupo el intervalo para el cual está definido y para los segundos las distintas categorías de la variable.

3. **Estadísticos que se desean estimar.** En el caso que solamente nos interese conocer el valor medio y el total, se pueden aplicar los métodos más sencillos como son: imputación al valor medio o mediano y en base a las proporciones pueden ser razonables. Sin embargo al aplicar estos métodos habrá problemas en la estimación de la varianza, debido a que se infraestima su valor real.

   En el caso en el que se requiera la distribución de frecuencias de la variable, la varianza y asociaciones entre las distintas variables, se deben emplear métodos mas elaborados y analizar el fichero de datos. El problema en este caso se incrementa cuando hay una elevada tasa de no-respuesta.

4. **Tasas de no-respuesta y exactitud necesaria.** No se debe abusar de los métodos de imputación y menos cuando tenemos una elevada tasa de no respuesta de la cual se desconoce el mecanismo. El problema no es tan grave en el caso en que se proporciona la correcta información sobre la precisión de las medidas estadísticas. En el artículo de Laaksonen (2000) se considera tasa de no-respuesta elevada cuando dicha tasa supera un tercio del total.

5. **Información auxiliar disponible.** La imputación puede mejorar al emplear información auxiliar disponible. En el caso de no disponer información auxiliar una técnica muy recomendada a aplicar es la imputación mediante el método hot deck aleatorio.

La tarea de la imputación varía en gran medida dependiendo del tamaño del conjunto de datos. Cuando se dispone de un fichero de datos pequeño es problemático en el caso de tener valores missing en unidades cruciales, al aplicar hot-deck aleatorio se pueden producir errores graves. Este caso se suele dar en muchas muestras económicas. En cambio cuando se posee un conjunto de datos de grandes dimensiones surgen menos problemas y se pueden aplicar distintos métodos de imputación.

La imputación se puede considerar como un proceso de varias etapas:

- **Paso 1:** El proceso de imputación empieza cuando se dispone de un fichero de datos con valores faltantes, que ha debido pasar anteriormente la fase de edición.

- **Paso 2:** Recopilar y validar para el proceso de imputación toda la información auxiliar que pueda ayudar en la imputación.

- **Paso 3:** Estudiar los distintos modelos de imputación para las variables que van a ser imputadas. Seleccionar la técnica de imputación a aplicar pudiendo ser: imputación univariante, en el caso de imputar una sola variable en cada momento ó imputación multivariante en el caso de imputar simultáneamente un conjunto de variables de la investigación estadística. En esta fase es interesante observar los patrones de no respuesta que aparecen en dicho estudio, y comprobar si hay gran número de registros que simultáneamente tienen no-respuesta en un conjunto de variables, en este caso puede ser interesante aplicar una imputación multivariante.

- **Paso 4:** Seleccionar varios métodos de imputación posibles. En esta fase según el tipo de la variable a imputar, información auxiliar disponible, tipo de no-respuesta,… se seleccionan los métodos apropiados para dicha variable. Es conveniente seleccionar más de uno para poder contrastar los resultados que se obtienen mediante los distintos métodos.

- **Paso 5:** Estimación puntual y varianza muestral para los distintos métodos de imputación empleados y su evaluación. El objetivo es obtener estimaciones con el mínimo sesgo y la mejor precisión.

- **Paso 6:** Tras estos se pasa a calcular la varianza de la imputación, la cual se puede calcular mediante diferentes técnicas. Durante los últimos años, se han presentado varios métodos para el cálculo de la estimación de la varianza de los datos imputados:
    - Imputación múltiple. Propuesto por Rubin (1987,1996).
    - Imputación de pesos fraccionada (Fractionally weighted imputation) basada en la imputación múltiple pero para la estimación de la varianza toma los beneficios de aplicar el método Jackknife propuesto por Rao y Shao (1992).
    - Analítica. Shao (1997) presenta algunos nuevos desarrollos referentes a algunos métodos de imputación para el cálculo de la varianza de los valores imputados.
- **Paso 7:** Resultados de la imputación.
    - Estimaciones puntuales y estimación final de la varianza.
    - Micro ficheros con valores reales e imputados.

3.4.- Criterios de cumplimiento por la Imputación

El proceso de imputación debe ser capaz de reproducir eficientemente un fichero de datos completo al cual se le pueda aplicar un análisis estadístico para datos completos. Con la finalidad de obtener unos resultados adecuados tras la imputación se deben calcular una serie de estadísticos que nos corroboren que estamos ante una imputación adecuada para el estudio en cuestión.

A continuación se proponen una serie de medidas que son deseables para obtener una buena imputación de datos, propuestas en el proyecto europeo de Edición e Imputación

de datos (EUREDIT). Para el caso en el cual se desean producir estimaciones agregadas los criterios 1. y 2. son irrelevantes.

1. **Precisión en la predicción**: El proceso de imputación debe preservar el valor real lo máximo posible, es decir, debe resultar un valor imputado que sea lo más cercano posible al valor real.

2. **Precisión en el ranking**: El proceso de imputación debe maximizar la preservación del orden en los valores imputados. Es decir, debe resultar una ordenación que relacione el valor imputado con el valor real o sea muy similar. Esta medida se refiere a variables numéricas o categóricas ordinales.

3. **Precisión en la distribución**: El proceso de imputación debe preservar la distribución de los valores reales. Es decir, las distribuciones marginales y de orden superior de los datos imputados debe ser esencialmente la misma que la correspondiente de los valores reales.

4. **Precisión en la estimación**: El proceso de imputación debe reproducir los momentos de órdenes menores de la distribución de los valores reales. En particular, debe producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.

5. **Imputación plausible:** El proceso de imputación debe conducir a valores imputados que sean plausibles. En particular, deben ser valores aceptables al aplicarles el proceso de edición.

Las medidas propuestas anteriormente dependen del tipo de variable que estemos considerando, según el tipo de las variables a imputar hay criterios que no hay que tener en cuenta.

Existen distintas medidas propuestas para los distintos tipos de variables (nominales, ordinales, continuas,..) que se pueden consultar en el artículo de EUREDIT **"Interm Report on Evaluation Criteria for Statistical Editing and Imputation"**. Principalmente las características que se desean obtener de la imputación realizada son: la conservación de los momentos de la distribución original y la semejanza entre los valores reales y los imputados asignados a cada uno de ellos.

3.5 Imputación múltiple

En las últimas décadas, se ha desarrollado un nuevo método en el área del análisis de datos incompletos: la imputación múltiple. Tras la publicación de los trabajos de Little y Rubin (1986-87) han aparecido otros muchos artículos estudiando esta técnica de imputación.

La imputación múltiple es una técnica en la que los valores perdidos son sustituidos por m >1 valores simulados. Consiste en la imputación de los casos perdidos a través de la estimación de un modelo aleatorio apropiado realizada m veces y, como resultado, se obtienen m archivos completos con los valores imputados. Posteriormente, se lleva a cabo el análisis estadístico ordinario con las m matrices de datos completas y se

combinan los resultados con una serie de fórmulas específicas proporcionadas por Little y Rubin (1987).

El objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no-respuesta parcial introduce en la estimación de parámetros. En el caso de imputación simple tiende a sobreestimar la precisión ya que no se tiene en cuenta la variabilidad de las componentes entre las distintas imputaciones realizadas. Para llevar a cabo la imputación múltiple de los valores perdidos, procederíamos del siguiente modo:

- En primer lugar se seleccionan las variables que se emplearán en el modelo de imputación. Es imprescindible que todas las variables que se van a utilizar conjuntamente en posteriores análisis se incluyan en dicho modelo, también se deben incluir todas aquellas variables que puedan ayudar a estimar los valores missing.
- En segundo lugar, se decide el número de imputaciones que se desea realizar. En general según se indica en la publicación de Rubin, entre 3 y 5 imputaciones son suficientes.
- Decidir el método de imputación a aplicar a los distintos ficheros de datos. Hay que tener en cuenta que esta fase es muy importante y se debe hacer un estudio del método a aplicar en función de las características de las variables a imputar, información auxiliar disponible, variables explicativas,… Para poder aplicar la imputación múltiple, el método seleccionado debe contener algún componente de imputación aleatoria. Con esta propiedad se asegura la posibilidad de obtener, para cada registro a imputar, modificaciones entre los valores imputados al completar los distintos ficheros de datos. Por ejemplo, no se va a poder aplicar la imputación múltiple en el caso de realizar métodos determinísticos, como pueden ser la imputación deductiva, al valor medio,…
- El siguiente paso será el de llevar a cabo los análisis estadísticos (univariantes, bivariantes o multivariantes) necesarios para la investigación. El análisis se realizará con las matrices generadas tras la imputación y los resultados se combinarán con las distintas fórmulas proporcionadas por Little y Rubin.

Observando las distintas matrices generadas tras la imputación múltiple se puede hacer una idea respecto a la precisión del método de imputación, si no se observan grandes variaciones entre los valores imputados de las distintas matrices se tiene una gran precisión de las estimaciones. Sin embargo hay técnicas estadísticas mas adecuadas para el estudio de la precisión de los estimadores.

El aspecto importante de la imputación múltiple, de la misma forma que en el resto de imputaciones, reside en la definición del modelo de imputación y en el método de imputación. Es fundamental que el modelo empleado en las estimaciones de los valores faltantes contenga las variables que se van a emplear posteriormente en los análisis estadísticos ordinarios, con el fin de preservar las relaciones entre las variables. Cuanto mejor sea el modelo respecto a la predicción, menor será la variación de los valores imputados y más precisos serán los estimadores posteriores. El método de estimación de los valores imputados varía de unas aplicaciones a otras, de modo que las propiedades también varían.

En general, la imputación múltiple es una de las soluciones más adecuadas al problema de no-respuesta parcial debido a su fácil aplicación y a la posibilidad de aplicar dicho método en distintas situaciones y ante diferentes tipos de variables.

En cuanto al software disponible para llevar a cabo estas técnicas, en la actualidad existen varias aplicaciones que permiten realizar la imputación múltiple con distintos tipos de matrices de datos. Entre las aplicaciones exclusivamente dedicadas a la imputación están los programas **AMELIA, MICE, NORM-CAT-MIX-PAN y SOLAS.**

Destacan los módulos de imputación múltiple incluidos recientemente en **SAS versiones 8.1 y 8.2.** También existen macros de SAS que realizan imputación múltiple: sirnorm, em_covar, mvn y macros de Paul Allinson. Existe además una aplicación SAS de imputación denominada **IVEvare.**