



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

<b>1</b>	<b>Introducción a las encuestas y formulación de objetivos y marcos</b>	<b>1</b>
1.1	Introducción a las encuestas y formulación de objetivos y marcos . . . . .	1
1.2	Esquema general de una encuesta . . . . .	2
1.3	Muestreo probabilístico . . . . .	6
1.4	Marco muestral . . . . .	7
1.5	Marco de áreas y otros marcos . . . . .	12
1.6	Población marco y población objetivo . . . . .	14
1.7	Planificación de una encuesta . . . . .	16
1.8	Diseño total de una encuesta . . . . .	19
	Bibliografía . . . . .	20
<b>2</b>	<b>Ideas básicas sobre estimación en muestreo probabilístico</b>	<b>1</b>
2.1	Ideas básicas sobre estimación en muestreo probabilístico . . . . .	1
2.2	Diseño muestral . . . . .	3
2.3	Probabilidades de inclusión . . . . .	4
2.4	La noción de estadístico . . . . .	6
2.5	Indicadores de pertenencia a la muestra . . . . .	7
2.6	Estimadores y sus propiedades básicas . . . . .	8
2.7	El estimador Horvitz-Thompson (estimador $\pi$ ) y sus propiedades . . . . .	13
2.8	Muestreo con reemplazamiento . . . . .	16
2.9	Efecto de diseño . . . . .	20
2.10	Intervalos de confianza . . . . .	21
	Bibliografía . . . . .	22
<b>3</b>	<b>Estimación insesgada en diseños muestrales sobre unidades elementales I</b>	<b>1</b>
3.1	Introducción . . . . .	1
3.2	Muestreo de Bernoulli . . . . .	2
3.2.1	Definición . . . . .	2
3.2.2	Estimadores, varianza y estimador de la varianza . . . . .	4
3.3	Muestreo aleatorio simple sin reemplazamiento . . . . .	8
3.3.1	Definición . . . . .	8
3.3.2	Estimadores, varianza y estimador de la varianza . . . . .	10
3.3.3	Estimación en dominios . . . . .	15
3.3.4	Comparación del muestreo aleatorio simple sin reemplazamiento y el muestreo de Bernoulli . . . . .	18
3.4	Muestreo aleatorio simple con reemplazamiento . . . . .	19
3.4.1	Definición . . . . .	19
3.4.2	Estimadores, varianza y estimador de la varianza . . . . .	20
3.4.3	Comparación del muestreo aleatorio simple sin y con reemplaza- miento . . . . .	22
	Bibliografía . . . . .	23

<b>4</b>	<b>Estimación insesgada en diseños muestrales sobre unidades elementales II.</b>	<b>1</b>
4.1	Introducción . . . . .	1
4.2	Muestreo sistemático: definición, estimadores, varianza del estimador .	2
4.2.1	El control del tamaño muestral . . . . .	7
4.3	La eficiencia del muestreo sistemático . . . . .	8
4.4	Problemática de la estimación de la varianza . . . . .	14
	Bibliografía . . . . .	18
<b>5</b>	<b>Estimación insesgada en diseños muestrales sobre unidades elementales III.</b>	<b>1</b>
5.1	Introducción . . . . .	1
5.2	Muestreo de Poisson . . . . .	2
5.2.1	Definición . . . . .	2
5.2.2	Estimadores, varianza y estimador de la varianza . . . . .	3
5.3	Muestreo con probabilidades proporcionales al tamaño . . . . .	5
5.3.1	Muestreo sin reemplazamiento . . . . .	6
5.3.2	Muestreo con reemplazamiento . . . . .	14
	Bibliografía . . . . .	17
<b>6</b>	<b>Estimación insesgada en diseños muestrales sobre unidades elementales IV</b>	<b>1</b>
6.1	Muestreo estratificado: definición, estimadores, varianza y estimador de la varianza . . . . .	1
6.1.1	Introducción y definición . . . . .	1
6.1.2	Estimadores, varianza y estimador de la varianza . . . . .	5
6.2	Afijación óptima . . . . .	10
6.2.1	Afijación en el caso de múltiples variables de estudio . . . . .	15
6.3	Otras afijaciones bajo muestreo aleatorio simple . . . . .	17
6.3.1	Afijación de Neyman . . . . .	17
6.3.2	Afijación óptima con información auxiliar . . . . .	17
6.3.3	Afijación proporcional . . . . .	18
6.3.4	Afijación proporcional al total de la variable y . . . . .	19
6.3.5	Afijación proporcional al total de una variable auxiliar . . . . .	19
6.4	Comparación de la precisión del estimador de Horvitz-Thompson en muestreo aleatorio estratificado según el tipo de afijación y el muestreo aleatorio simple . . . . .	20
	Bibliografía . . . . .	25
<b>7</b>	<b>Estimación insesgada en diseños muestrales por conglomerados I.</b>	<b>1</b>
7.1	Definición, estimadores, varianza y estimador de la varianza . . . . .	1
7.1.1	Introducción y definiciones . . . . .	1
7.1.2	Estimadores, varianza y estimador de la varianza . . . . .	4
7.1.3	Muestreo por conglomerados aleatorio simple . . . . .	9
	Bibliografía . . . . .	17
<b>8</b>	<b>Métodos y gestión de la recogida de datos.</b>	<b>1</b>
8.1	Introducción a los métodos de la recogida de datos . . . . .	1
8.2	Métodos de recogida de datos básicos . . . . .	2
8.3	Recogida de datos asistida por ordenador . . . . .	4

8.4	Otros métodos de recogida . . . . .	6
8.4.1	Observación directa . . . . .	6
8.4.2	Intercambio electrónico de datos (EDI, <i>Electronic Data Interchange</i> ) . . . . .	6
8.4.3	Datos administrativos . . . . .	8
8.4.4	Modos combinados (Mixed-modes) . . . . .	9
8.5	Introducción a la gestión de la recogida de datos . . . . .	11
8.6	Implementación de la encuesta . . . . .	12
8.7	Gestión activa del trabajo de campo . . . . .	14
8.8	Paradatos . . . . .	16
8.9	Monitorización de la calidad de la respuesta . . . . .	17
8.10	Monitorización del proceso de producción de una encuesta . . . . .	18
8.11	Evaluación de la encuesta y el informe de calidad . . . . .	19
	Bibliografía . . . . .	20
<b>9</b>	<b>Introducción a la depuración e imputación de datos estadísticos en el proceso estadístico.</b>	<b>1</b>
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) . . . . .	3
9.2.1	Tipos de errores . . . . .	4
9.2.2	Tipos de datos <i>missing</i> . . . . .	7
9.2.3	Reglas de depuración . . . . .	8
9.3	Métodos básicos para la depuración e imputación de datos estadísticos . . . . .	11
9.3.1	Depuración durante la fase de recogida de datos . . . . .	12
9.3.2	Métodos modernos de depuración . . . . .	12
9.3.3	Métodos de imputación . . . . .	13
9.4	Estrategia de depuración e imputación . . . . .	17
	Bibliografía . . . . .	19
<b>10</b>	<b>Introducción a la estimación en presencia de falta de respuesta.</b>	<b>1</b>
10.1	Introducción a la estimación en presencia de falta de respuesta . . . . .	1
10.2	Errores debidos al muestreo y a la falta de respuesta . . . . .	3
10.3	Error cuadrático medio y sus componentes bajo falta de respuesta . . . . .	5
10.4	Estimadores simples y sus sesgos debidos a la falta de respuesta . . . . .	6
10.4.1	Vector auxiliar . . . . .	7
10.4.2	Clasificación unidireccional . . . . .	7
10.4.3	Una sola variable auxiliar cuantitativa . . . . .	11
10.4.4	Clasificación unidireccional combinada con un variable cuantitativa . . . . .	13
10.4.5	Clasificación bidireccional . . . . .	15
10.4.6	Estimadores simples y su sesgo por falta de respuesta . . . . .	16
	Bibliografía . . . . .	19
<b>11</b>	<b>Imputación.</b>	<b>1</b>
11.1	Introducción . . . . .	1
11.2	¿Qué es la imputación? . . . . .	4
11.3	Terminología . . . . .	5
11.4	Múltiples variables de estudio . . . . .	8

11.5	El enfoque de imputación completa . . . . .	9
11.6	El enfoque combinado . . . . .	10
11.7	El enfoque de reponderación completa . . . . .	11
11.8	Imputación por reglas estadísticas . . . . .	13
11.8.1	Imputación por regresión . . . . .	14
11.8.2	Imputación por el vecino más cercano . . . . .	15
11.8.3	Imputación <i>hot deck</i> . . . . .	15
11.8.4	Grupos de imputación . . . . .	16
11.8.5	Introducción de un residuo seleccionado aleatoriamente . . . . .	17
11.9	Imputación por juicio del experto y por datos históricos . . . . .	18
	Bibliografía . . . . .	20
<b>12</b>	<b>Control del secreto estadístico.</b>	<b>1</b>
12.1	Conceptos y definiciones: Control del secreto estadístico, datos tabulares, microdatos, riesgo y utilidad . . . . .	1
12.2	Un enfoque el control del secreto estadístico: por qué la protección de la confidencialidad es importante, características clave y usos de los datos, riesgos contra los que la protección es necesaria, métodos de control del secreto, implementación . . . . .	5
12.2.1	Aplicación a resultados en tablas . . . . .	5
12.2.2	Aplicación a microdatos . . . . .	10
12.3	Conclusiones . . . . .	18
	Bibliografía . . . . .	18
<b>13</b>	<b>Difusión de datos: Presentación de estadísticas.</b>	<b>1</b>
13.1	Introducción . . . . .	1
13.2	Transmitir el mensaje . . . . .	1
13.3	Visualización de las estadísticas . . . . .	3
13.4	Tablas . . . . .	6
13.5	Gráficos . . . . .	8
13.6	Mapas . . . . .	14
13.7	Técnicas de visualización emergentes . . . . .	18
13.8	Cuestiones de accesibilidad . . . . .	19
	Bibliografía . . . . .	20
<b>14</b>	<b><i>Record linkage</i>.</b>	<b>1</b>
14.1	Introducción . . . . .	1
14.2	Los datos administrativos en la estadística oficial . . . . .	3
14.3	Visión conjunta de los métodos . . . . .	7
14.3.1	El modelo de <i>record linkage</i> de Fellegi-Sunter . . . . .	7
14.3.2	Parámetros de aprendizaje . . . . .	8
14.3.3	Comparadores de cadenas . . . . .	12
14.3.4	Datos de entrenamiento . . . . .	13
14.4	Preparación de los datos . . . . .	13
14.4.1	Descripción de un proyecto de <i>matching</i> . . . . .	13
14.4.2	Preparación inicial de ficheros . . . . .	14
14.4.3	Estandarización y análisis sintáctico de nombres . . . . .	16

14.4.4 Estandarización y análisis sintáctico de direcciones . . . . .	17
14.4.5 Estandarización y normalización de registros administrativos . .	17
14.4.6 Resumen sobre el preprocesamiento . . . . .	21
14.5 Caso práctico con registros administrativos . . . . .	21
Bibliografía . . . . .	22
<b>15 Metadatos de la producción Estadística. I.</b>	<b>1</b>
15.1 Introducción . . . . .	1
15.2 El modelo . . . . .	2
15.2.1 La estructura . . . . .	3
15.2.2 Aplicabilidad . . . . .	4
15.2.3 El uso del GSBPM . . . . .	5
15.3 Relaciones con otros modelos y estándares . . . . .	6
15.3.1 GAMS0 . . . . .	6
15.3.2 GSIM . . . . .	7
15.4 Niveles 1 y 2 del GSBPM . . . . .	8
15.5 Descripciones de fases y subprocesos (fases 1 a 3) . . . . .	10
Bibliografía . . . . .	17
<b>16 Metadatos de la producción Estadística. II.</b>	<b>1</b>
16.1 Descripciones de fases y subprocesos (fases 4 a 8) . . . . .	1
16.2 Procesos generales ( <i>overarching processes</i> ) . . . . .	12
16.3 Otros usos del GSBPM . . . . .	16
16.4 Data Documentation Initiative (DDI), SDMX y comparación con el GSBPM	18
Bibliografía . . . . .	20
<b>17 Metadatos de la producción Estadística. III.</b>	<b>1</b>
17.1 Introducción genérica al GSIM . . . . .	1
17.2 Introducción al Documento de Comunicación del GSIM . . . . .	2
17.3 Alcance . . . . .	3
17.4 ¿Qué es el GSIM? . . . . .	3
17.5 Beneficios del GSIM para la organización como un todo . . . . .	7
17.6 Relación con otros modelos ModernStats: GSIM y GSBPM . . . . .	9
17.7 ¿Qué implica para el estadístico? . . . . .	11
17.8 SDMX, DDI y otros estándares . . . . .	15
Bibliografía . . . . .	20
<b>18 La calidad en la estadística oficial y el Código de Buenas Prácticas de las Estadísticas Europeas.</b>	<b>1</b>
18.1 El concepto de calidad en la estadística oficial . . . . .	2
18.2 El Código de Buenas Prácticas de las Estadísticas Europeas . . . . .	4
18.3 El marco de garantía de la calidad del Sistema Estadístico Europeo . . .	14
18.4 La calidad en los productos y en los procesos estadísticos . . . . .	18
18.5 Sistemas de evaluación global de la calidad: auditorías, autoevaluación y revisiones por homólogos en las oficinas de Estadística . . . . .	21
Bibliografía . . . . .	23





## Tema 1

**Introducción a las encuestas y formulación de objetivos y marcos. Esquema general de una encuesta. Muestreo probabilístico. Marco muestral. Marco de áreas y otros marcos. Población marco y población objetivo. Planificación de una encuesta. Diseño global de una encuesta.**

Este tema está elaborado como una adaptación casi literal en español del capítulo 1 de la siguiente bibliografía:

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

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.

### 1.1 Introducción a las encuestas y formulación de objetivos y marcos

La necesidad de información estadística parece interminable en la sociedad moderna en que vivimos. En particular, se recogen datos de forma regular para satisfacer la necesidad de información sobre conjuntos específicos de elementos, llamados poblaciones finitas. Por ejemplo, nuestro objetivo puede ser obtener información sobre los hogares en una ciudad y sus comportamientos en materia de gastos, empresas en una determinada industria y sus beneficios, las personas de un país y su situación laboral, o las granjas de una región y su producción de cereales.

Una de las formas más importantes de recogida de datos en la producción estadística oficial para satisfacer estas necesidades es una encuesta muestral, es decir, una investigación parcial de la población finita a través de una encuesta. Una encuesta muestral cuesta menos que un censo, es más rápida y puede ser, incluso, más acurada<sup>1</sup> que los censos.

---

<sup>1</sup>Empleamos el término *acurado* como traducción literal de *accurate* pensando en el indicador del error cuadrático medio, por distinción al término *preciso*, cuyo uso restringimos para referirnos solo a la varianza de estimador (véase más abajo).

A lo largo del siglo XX el muestreo con encuestas<sup>2</sup> ha evolucionado hacia un conjunto de teorías, métodos y operaciones usadas diariamente en todo el mundo.

En muchos países, se constituye legalmente un instituto nacional de estadística con el fin de proporcionar información estadística sobre la situación del país. Las encuestas son una parte importante de esta actividad. Por ejemplo, en España, el Instituto Nacional de Estadística (INE) se rige, básicamente, por la Ley 12/1989, de 9 de mayo, de la Función Estadística Pública (LFEP)<sup>3</sup>, que regula la actividad estadística para fines estatales, que es competencia exclusiva del Estado.

Por tanto, los INEs producen regularmente estadísticas sobre características y actividades nacionales importantes, incluyendo la demografía (distribución por edad y sexo, fertilidad, mortalidad), agricultura (distribución de las cosechas), población activa (empleo), salud y condiciones de vida, industria y comercio. Gran parte de la teoría básica de muestreo se desarrolló en oficinas de estadística.

En las universidades, el muestreo es ampliamente utilizado, especialmente en sociología e investigación de la opinión pública, y también en economía, ciencias políticas y psicología. El muestreo ha crecido mucho y es un enfoque hoy día aceptado universalmente como forma de obtener información. Todos los años se dedican muchos recursos a realizar encuestas.

Los medios de comunicación proporcionan al público resultados de encuestas nuevas o periódicas. Y es ampliamente aceptado que una muestra puede proporcionar una imagen acurada de una población más grande<sup>4</sup>; por ejemplo, una muestra bien seleccionada de unas miles de personas puede describir una población de varios millones. Sin embargo, reunir los datos es muy costoso. Por tanto, por razones de efectividad de los costes, es imprescindible usar los mejores métodos disponibles para diseñar las muestras y para el cálculo de estimaciones, utilizar la información auxiliar, etcétera.

## 1.2 Esquema general de una encuesta

Para empezar, necesitamos un esquema general de una encuesta y una terminología básica. Los términos *encuesta* (*survey*) y *muestreo con encuestas* (*survey sampling*) se usan para denotar investigaciones estadísticas que tienen las siguientes características metodológicas:

---

<sup>2</sup>Traducimos *survey sampling* como *muestreo con encuestas* para hacer explícitos los dos elementos fundamentales de este modo de producir información (muestras y encuestas), que distinguirlo de otros (censos, registros administrativos, nuevas fuentes de datos digitales).

<sup>3</sup><https://www.boe.es/buscar/doc.php?id=BOE-A-1989-10767>

<sup>4</sup>No obstante, desde la propuesta original a finales del s. XIX de emplear muestras pasaron alrededor de 40 años hasta que empezaron a emplearse de manera rutinaria.

- i. Una encuesta hace referencia a un conjunto finito de *elementos* llamado *población finita*. Existe una regla de enumeración que define de forma unívoca a los elementos de la población. El objetivo de una encuesta es proporcionar información sobre la población finita en cuestión o sobre subpoblaciones de especial interés, por ejemplo, 'hombres' y 'mujeres' pueden ser dos subpoblaciones de 'todas las personas'. Estas subpoblaciones se denominan *dominios de estudio* o simplemente *dominios*.
- ii. Se asocia el valor de una o más *variables de estudio* (a veces también *variables objetivo* (*target variables*)) con cada elemento de la población. El objetivo de una encuesta es obtener información sobre *características de la población*, *parámetros*, *agregados* o *indicadores* desconocidos. Los parámetros son funciones de los valores de las variables de estudio. Son medidas cuantitativas desconocidas de interés, por ejemplo, los ingresos totales, los ingresos medios, la producción total, el número de desempleados, tanto para la población completa como usualmente para dominios específicos.
- iii. En la mayoría de las encuestas, el acceso y la observación de elementos individuales de la población se establece mediante un *marco de muestreo*, un recurso que asocia los elementos de la población con las *unidades muestrales* en el marco.
- iv. A partir de la población, se selecciona una *muestra* (es decir, un subconjunto) de elementos. Esto se puede llevar a cabo seleccionando unidades del marco. Una muestra será una *muestra probabilística* si se obtiene utilizando un mecanismo aleatorio como se verá en la Sección 1.3.
- v. Los elementos de la muestra son *observados*, es decir, para cada elemento de la muestra, las variables de estudio son *medidas* y sus valores son *grabados*. La medición se ajusta a un *plan de medición* bien definido, especificado en términos de instrumentos de medida, una o más operaciones de medida, el orden entre estas y las condiciones bajo las cuales se llevan a cabo.
- vi. Los valores de las variables grabados se usan para calcular *estimaciones (puntuales)* de los parámetros de interés de la población finita (totales, medias, medianas, proporciones, coeficientes de regresión, etc.). También se calculan estimaciones de la precisión de las estimaciones. Las estimaciones son finalmente publicadas.

En una encuesta por muestreo, la observación (medición) se limita a un subconjunto de la población. El tipo de encuestas en las que se observa/mide toda la población se llama *censo*.

*Ejemplo 1.* Las encuestas de población activa se llevan a cabo en muchos países. Estas encuestas tienen como objetivo responder preguntas como: ¿Cuántas personas activas hay en el país y en cada una de sus regiones? ¿Qué proporción de éstas están desempleadas? En este caso, algunos de los conceptos clave son los siguientes. *Población:* Todas las personas del país con ciertas excepciones (como menores de 16 años, personas ingresadas en instituciones). *Dominios de interés:* Grupos por edad y sexo de la población, grupos por ocupación y regiones del país. *Variables:* Cada persona, en el momento de la

encuesta, se puede clasificar en (a) perteneciente a la población activa o no, y (b) empleada o no. Por tanto, hay una variable de interés que toma el valor 'uno' si la persona pertenece a la población activa y 'cero', en caso contrario. Para medir el desempleo, se define una segunda variable de interés que toma el valor 'uno' si una persona está desempleada, 'cero', en caso contrario. Son esenciales las definiciones precisas. Si el motivo es estimar el desempleo en un mes determinado y una persona entrevistada indica que ha trabajado una semana durante ese mes, pero está desempleada el día de la entrevista, debe haber una regla precisa que indica si esa persona está desempleada o no. *Características de interés de la población:* Número de personas activas/ocupadas/paradas/inactivas, proporción de ocupados/parados en la población activa. *Muestra:* Se selecciona una muestra de personas de la población de la manera más eficiente teniendo en cuenta los recursos existentes. *Observaciones/mediciones:* Un entrevistador visita a cada persona incluida en la muestra, le pregunta las cuestiones incluidas en un cuestionario estandarizado y graba las respuestas. *Procesamiento de datos y estimación:* Los datos grabados son depurados, es decir, se preparan para la fase de estimación; se tienen en cuenta las reglas para la falta de respuesta; se calculan las estimaciones de las características de la población. Se calculan indicadores sobre la precisión de las estimaciones. Se publican los resultados. ■

*Ejemplo 2.* Consideremos una encuesta a hogares cuyo objetivo es obtener información sobre los gastos planificados por los hogares para el año para un bien específico. En este caso, algunos de los conceptos clave son los siguientes. *Población:* Todos los hogares del país. *Variables:* Gasto planificado en euros para bienes específicos, como coches, neveras, etc. *Características de interés de la población:* Gasto total planificado por hogar para los bienes específicos. *Muestra:* Se obtiene una muestra de hogares seleccionando, inicialmente, una muestra de áreas geográficas y, a continuación, una submuestra de hogares en las áreas seleccionadas. *Observaciones/mediciones:* Cada hogar seleccionado rellena un cuestionario (electrónico o en papel). La mayoría de los hogares responden al cuestionario. Las que no responden son recontactadas por teléfono o en persona. *Procesamiento de datos y estimación:* Los datos son depurados. Se calculan las estimaciones y la precisión teniendo en cuenta el diseño en dos etapas. ■

Esta visión de la producción como un proceso en fases y etapas es fundamental para la modernización e industrialización de la producción estadística oficial y está íntimamente relacionada con los estándares internacionales como el GSBPM (UNECE 2019b)<sup>5</sup> y el GSIM (UNECE 2019a)<sup>6</sup>.

Las características metodológicas (i) a (vi) identificadas anteriormente dan lugar a varios comentarios.

1. La complejidad de una encuesta puede variar mucho, dependiendo del tamaño de la población y de los medios para acceder a la población. Encuestar a los socios de una asociación, los hospitales de una región o los residentes en un pequeño pueblo puede ser relativamente sencillo. En el otro extremo están las encuestas

---

<sup>5</sup>Véanse los temas 15 y 16

<sup>6</sup>Véase tema 17

complejas a nivel nacional, con una población de millones de personas residiendo en un territorio grande; este tipo de encuestas son llevadas a cabo por los INEs y requieren muchos recursos administrativos y económicos.

2. Aunque las encuestas suponen observaciones de elementos individuales de la población, el motivo de una encuesta no es usar estos datos para tomar decisiones sobre elementos individuales, sino para obtener estadísticas generales sobre la población o subgrupos específicos.
3. En la misma encuesta a menudo hay muchas variables de estudio y muchos dominios de interés. El número de características a estimar puede ser grande, cientos o miles.
4. Los parámetros de poblaciones finitas son medidas cuantitativas de varios aspectos de la población. Antes de la encuesta estos parámetros son desconocidos. Hay distintos tipos de parámetros: el total, la media, la mediana de la variable de estudio, el coeficiente de variación entre dos variables, etcétera. El valor exacto del parámetro puede ser obtenido en casos especiales, si se realiza un censo, no hay errores de medida y sin falta de respuesta. Un censo no significa automáticamente 'estimación sin error'.
5. Una muestra es cualquier subconjunto de la población. Puede ser seleccionada con un mecanismo aleatorio o no. Un ejemplo sencillo de un esquema de diseño aleatorio es uno que da a cada muestra de tamaño fijo la misma probabilidad de selección (esto es un *muestreo aleatorio simple sin reemplazamiento*). En la práctica, los esquemas de selección son más complejos. El muestreo probabilístico ha demostrado a lo largo de los años ser un instrumento acurado y se ha convertido en la herramienta fundamental para hacer inferencia a partir de una muestra en la producción estadística oficial. El muestreo también puede ser no probabilístico, por ejemplo, seleccionando unidades muestrales mediante el criterio de un experto. En estos casos, el control de la acuracidad (sesgo y varianza) es prácticamente imposible o depende de hipótesis de difícil comprobación.
6. Medir de forma correcta y grabar la información necesaria para todos los elementos de la muestra puede ser difícil o imposible. Se pueden obtener respuestas falsas o erróneas. Para algunas unidades de la muestra puede no ser posible obtener los datos por ser imposible el contacto o por negativa a responder. Estos llamados *errores ajenos al muestreo* pueden ser grandes y afectar muy negativamente a la acuracidad de la operación estadística.
7. Los avances en la informática han hecho posible producir un gran número de estadísticas oficiales a partir de datos administrativos. Se pueden usar muchos ficheros. Por ejemplo, se cruzan los elementos de dos registros completos de población. Los ficheros cruzados proporcionan una base mayor para la producción de estadísticas. También se puede combinar la información de una encuesta muestral con información de uno o más registros administrativos. Los datos administrativos pueden entonces servir como información auxiliar para fortalecer las estimaciones.

## 1.3 Muestreo probabilístico

El *muestreo probabilístico* es un enfoque de la selección de muestras que satisface determinadas condiciones, las cuales, para el caso de selección directa de elementos de la población, se describen a continuación:

1. Podemos definir el conjunto de muestras  $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$  que se pueden obtener con el proceso de muestreo.
2. Existe una probabilidad de selección conocida  $p(s)$  asociada a cada muestra posible  $s \in \mathcal{S}$ .
3. El procedimiento otorga a cada elemento en la población una probabilidad no nula de selección.
4. Se selecciona una muestra mediante un mecanismo aleatorio bajo el cual cada posible muestra  $s \in \mathcal{S}$  recibe exactamente una probabilidad  $p(s)$ .

Una muestra obtenida bajo estas cuatro condiciones se llama *muestra probabilística*.

Si la encuesta funciona correctamente sin alteraciones, podemos medir cada elemento de la muestra seleccionada y obtener los valores reales observados para las variables de estudio. Asumimos que existe una fórmula para calcular una estimación de cada parámetro de interés. Los datos de la muestra se incluyen en la fórmula, dando lugar, para cada muestra posible, a una estimación única.

La función  $p(s)$  define una distribución de probabilidad sobre  $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$ . Se llama *diseño de muestreo*, *diseño muestral* o, simplemente, *diseño*. Se verá una definición más rigurosa en la Sección 2.2.

La probabilidad a la que se refiere el punto 3 se llama *probabilidad de inclusión* del elemento (a veces también *probabilidad de inclusión de primer orden*). Bajo un diseño muestral probabilístico cada elemento de la población tiene una probabilidad de inclusión estrictamente positiva. Este requisito es muy exigente, pero juega un papel muy importante en el enfoque del muestreo probabilístico. En la práctica algunas veces hay razones irrefutables para que no se verifique de forma estricta este requisito. El muestreo por *cut-off* (que se utiliza en el INE en algunas encuestas coyunturales dirigidas a establecimientos industriales, por ejemplo) es una técnica en la que determinados elementos son excluidos de forma deliberada de la selección. En ese caso, las conclusiones válidas se limitan a la parte de la población que puede ser muestreada.

La aleatorización a la que se refiere el punto 4 se lleva a cabo normalmente mediante la implementación de un algoritmo. Un tipo común de algoritmos es aquel en que se realiza un experimento aleatorizado dando lugar a la inclusión o exclusión de cada elemento del marco en la muestra. Dos referencias básicas sobre algoritmos muestrales para la selección de muestras probabilísticas son (Brewer y Hanif 1983; Tillé 2006).

El muestreo a menudo se realiza en dos o más etapas. En primera etapa se seleccionan conglomerados de elementos. Esto puede venir seguido de una o más etapas de submuestreo; los propios elementos son muestreados en la última etapa. Para tener un diseño de muestreo probabilístico en estos casos se deben verificar las condiciones 1 a 4 anteriores en cada etapa. El procedimiento en su conjunto debe proporcionar a cada elemento de la población una probabilidad de inclusión estrictamente positiva.

El muestreo probabilístico ha evolucionado hacia un enfoque científico importante. Las dos principales razones para la selección aleatorizada (muestras probabilísticas) son (1) la eliminación de los sesgos de selección y (2) las muestras seleccionadas aleatoriamente son 'objetivas' y aceptables para el público. Estas dos mismas razones suponen un reto en la actualidad para la incorporación de nuevas fuentes de datos (como Big Data) en la producción estadística oficial.

## 1.4 Marco muestral

El *marco* o *marco muestral* es cualquier material o recurso usado para obtener acceso a la población finita de interés. Con la ayuda del marco debe ser posible (1) identificar y seleccionar una muestra de forma que respete un diseño muestral probabilístico y (2) establecer contacto con los elementos seleccionados (por teléfono, correo, dirección postal, etc.). La siguiente definición es de (Lessler y Kalsbeek 1992):

### Definición 1

Un *marco muestral* consiste en materiales, procedimientos y recursos que identifican, distinguen y permiten el acceso a los elementos de la población objetivo. Se compone de un conjunto finito de unidades al que se aplica el diseño muestral probabilístico [o no probabilístico, en sentido más general]. Incluye también información auxiliar (medidas de tamaño, información demográfica) usadas para (1) técnicas muestrales especiales, como la estratificación o la selección muestral proporcional al tamaño o (2) técnicas de estimación especiales, como la estimación de razón o de regresión.

Los marcos muestrales son listas o procedimientos para identificar todos los elementos de la población objetivo. Pueden ser mapas de áreas donde pueden encontrarse a los elementos de la población (véase la Sección 1.5). En su concepción más simple, un marco muestral constituye una sencilla lista de elementos de la población. Existen poblaciones para las que tales listas verdaderamente existen y están disponibles, como los establecimientos industriales de un país, los hospitales, las escuelas y otro tipo de instituciones. En las oficinas de estadística existen también registros de personas y/o direcciones postales de viviendas que pueden servir como marcos de personas.

Como en la definición anterior, llamamos *elementos*<sup>7</sup> a las entidades que forman la

<sup>7</sup>A veces también *unidades estadísticas de la población* o, a secas, *unidades estadísticas*.



población y *unidades muestrales* (a veces solo *unidades*) a las entidades del marco. Es conveniente también distinguir las *unidades informantes* (Statistics Canada 2010). Por ejemplo, en una encuesta en la que quieran analizarse menores de edad, los elementos de la población serían las personas menores de 18 años. Ahora bien, para acceder a ellos a menudo se selecciona una muestra de un marco de viviendas, que serían las unidades muestrales. Por último, al ser menores, es posible que la información sea proporcionada por un adulto (padre/madre/tutor legal), que sería la unidad informante. En muchas operaciones estadísticas, los tres tipos de unidades coinciden.

*Ejemplo 3.* El Padrón Municipal <sup>8</sup> es un marco que contiene información sobre todos los vecinos de los municipios de España. Este marco contiene, para cada individuo, información sobre variables como la fecha y el lugar de nacimiento, el sexo, la nacionalidad o el domicilio habitual. Faltan algunas personas, e incluye algunas que realmente no pertenecen a él, pero es un buen marco muestral. Una característica muy interesante es que proporciona acceso directo a la población de España. A menudo se usa un muestreo estratificado a partir de este marco para las encuestas dirigidas a personas llevadas a cabo por el INE. Se puede contactar fácilmente con los elementos muestreados (individuos). ■

*Ejemplo 4.* El Directorio Central de Empresas (DIRCE) <sup>9</sup> es el marco muestral usado en el INE para las encuestas a empresas. Es un marco bastante complejo y está basado en la información de varias fuentes. Por un lado, utiliza información de registros administrativos, como el *Impuesto sobre el Valor Añadido*, el *Impuesto de Sociedades* y el *Impuesto sobre la Renta de las Personas Físicas* de la Agencia Estatal de Administración Tributaria, el *Registro de Cuentas de Cotización* y el *Registro de Trabajadores Activos en Cuenta Propia* de la Seguridad Social, los movimientos del Registro Mercantil y también información de las encuestas estructurales y coyunturales de empresas realizadas por el INE. Es necesaria la actualización continua para registrar los 'nacimientos' (nuevas empresas que inician su actividad), 'muertes' (finalización de la actividad de la empresa) y cambios en la clasificación basados en el tamaño, la actividad o su ubicación geográfica. ■

Usaremos el término *muestreo directo de elementos* para denotar la selección muestral de un marco que identifica directamente a los elementos individuales de la población de interés. Es decir, las unidades del marco son objetos del mismo tipo que aquellos que queremos medir y observar. Una selección de elementos puede tener lugar directamente del marco. De forma ideal, el conjunto de elementos identificados en el marco coincide con el conjunto de elementos en la población de interés.

Por ejemplo, si la población de interés son los individuos residentes en España, podemos llevar a cabo un muestreo directo de elementos a partir del Padrón Municipal

---

<sup>8</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736177011&menu=resultados&idp=1254734710990](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177011&menu=resultados&idp=1254734710990)

<sup>9</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550)



indicado en el Ejemplo 3. Aquí, la unidad muestral coincide con el elemento, que es el individuo. (Los dos conjuntos realmente no son exactamente iguales, pero las diferencias son pequeñas). El marco del Ejemplo 4 se puede usar para el muestreo directo de elementos con el objetivo de estudiar la población de empresas en España; en este caso, las unidades muestrales coinciden con los elementos, que son las empresas.

Un marco muestral debería incluir idealmente los siguientes elementos ([Statistics Canada 2010](#)):

1. Datos de identificación. Son las variables del marco muestral que sirven para identificar de forma única cada unidad muestral, por ejemplo, nombre, apellidos, edad, sexo y un número identificador. En el Padrón Municipal el identificador es el número de Documento Nacional de Identidad o, tratándose de extranjeros, del documento que lo sustituya y, en el caso del DIRCE, es el Número de Identificación Fiscal.
2. Datos de contacto. Son las variables necesarias para localizar a las unidades muestrales durante la recogida, por ejemplo, la dirección postal o el número de teléfono.
3. Datos de clasificación. Son variables útiles tanto para la selección muestral como para la estimación. Por ejemplo, si la gente que viva en apartamentos tiene que ser encuestada de forma diferente a la que vive en casa, entonces el marco debe clasificar los distintos tipos de viviendas (p.ej. apartamentos, casas independientes, etc.). Los datos de clasificación pueden también incluir una medida de tamaño que se pueda usar en el muestreo, por ejemplo, el número de empleados trabajando en una empresa o el número de hectáreas de una explotación agrícola. Otros ejemplos de clasificaciones son clasificaciones geográficas (p.ej. la provincia, el municipio, la sección censal), la clasificación nacional de actividades económicas, etc.
4. Datos de mantenimiento. Datos necesarios si la encuesta se repetirá en otro momento posterior, por ejemplo, fechas de incorporación o de cambios en los datos del marco.
5. Datos de cruce o de enlace. Son variables que se usan para enlazar las unidades muestrales con una fuente de datos más actualizada, por ejemplo, para actualizar el marco de la encuesta.

Se simplifica el procedimiento de selección de la muestra si se verifica lo siguiente:

6. El marco está organizado de forma sistemática, por ejemplo, las unidades están ordenadas por tamaño o por situación geográfica.

Otra información a menudo está disponible en el marco y a menudo mejorará las estimaciones. Las siguientes características son deseables:

7. El marco contiene información adicional para cada unidad; esta información puede usarse para mejorar la eficiencia, como en la estratificación, o para construir estimadores que involucren variables auxiliares.

8. Cuando es necesaria una estimación por dominios (subpoblaciones), el marco especifica el dominio al cual pertenece cada unidad.

Otras propiedades deseables implican la relación entre las unidades en el marco y los elementos de la población:

9. Cada elemento en la población de interés está presente una única vez en el marco.
10. Ningún elemento que no esté en la población de interés estará en el marco.

Estas dos características simplificarán muchos procedimientos de selección y de estimación.

12. Cada elemento de la población de interés está presente en el marco.

La última propiedad es particularmente importante, porque si no se verifica el marco no proporciona acceso al total de la población de interés. En tal caso, ni siquiera la observación de todos los elementos en el marco haría posible calcular el verdadero valor del parámetro de la población finita de interés.

En la práctica, un marco a menudo toma la forma de un fichero de datos. Como mínimo, es un fichero con un elemento identificador  $k$  que va desde 1 hasta  $N_F$ . Puede contener otra información, como la indicada en los puntos 7 y 8. Podemos especificar todo lo que está disponible en el marco acerca de la  $k$ -ésima unidad muestral con un vector<sup>10</sup>  $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{jk}, \dots, x_{qk})^t$ . Aquí,  $x_{jk}$  es el valor de la  $j$ -ésima variable en la  $k$ -ésima unidad muestral. El valor  $x_{jk}$  puede ser cuantitativo (por ejemplo, la cifra de negocios del individuo  $k$ ) o cualitativa (por ejemplo, la dirección del individuo  $k$ ). El marco se puede ver como una matriz con  $N_F$  filas (registros) y con cada fila se asocian  $q + 1$  entradas de datos (campos); una entrada por identificador y  $q$  entradas para las componentes del vector  $\mathbf{x}_k$ , como a continuación:

Identificador	Vector conocido
1	$\mathbf{x}_1^t$
2	$\mathbf{x}_2^t$
$\vdots$	$\vdots$
k	$\mathbf{x}_k^t$
$\vdots$	$\vdots$
$N_F$	$\mathbf{x}_{N_F}^t$

En resumen, las características básicas que un marco muestral debe tener idealmente son (Burg y col. 2019):

- el marco debe estar disponible en formato digital;
- el marco tiene por objeto representar la población objetivo tan acuradamente como sea posible;

<sup>10</sup>Todos los vectores se entenderán como vectores columna; de ahí, el uso de  $^t$  para indicar transposición cuando se escribe el vector en fila.

- el marco contiene las unidades muestrales básicas correspondientes a los elementos de la población objetivo y asigna a cada unidad muestral un identificador único;
- el marco incluye variables de enlace (*linking variables*), que permiten conectar las unidades muestrales básicas con registros externos;
- el marco está enriquecido con variables auxiliares, permitiendo un mejor uso (al menos con las variables de contacto);
- Si existen unidades muestrales compuestas de unidades muestrales básicas (por ejemplo, hogares a partir de personas), el enlace entre ambos tipos de unidades está incluido en el marco.

Por último, dada la proliferación del uso de registros administrativos en las oficinas de estadística, es importante reseñar las diferencias entre estos y los marcos muestrales. Un registro es un conjunto completo escrito de registros que contiene entradas de elementos y detalles sobre un conjunto particular de objetos<sup>11</sup> Debe distinguirse entre registros *administrativos* y registros *estadísticos* (véase la Figura 1.1).

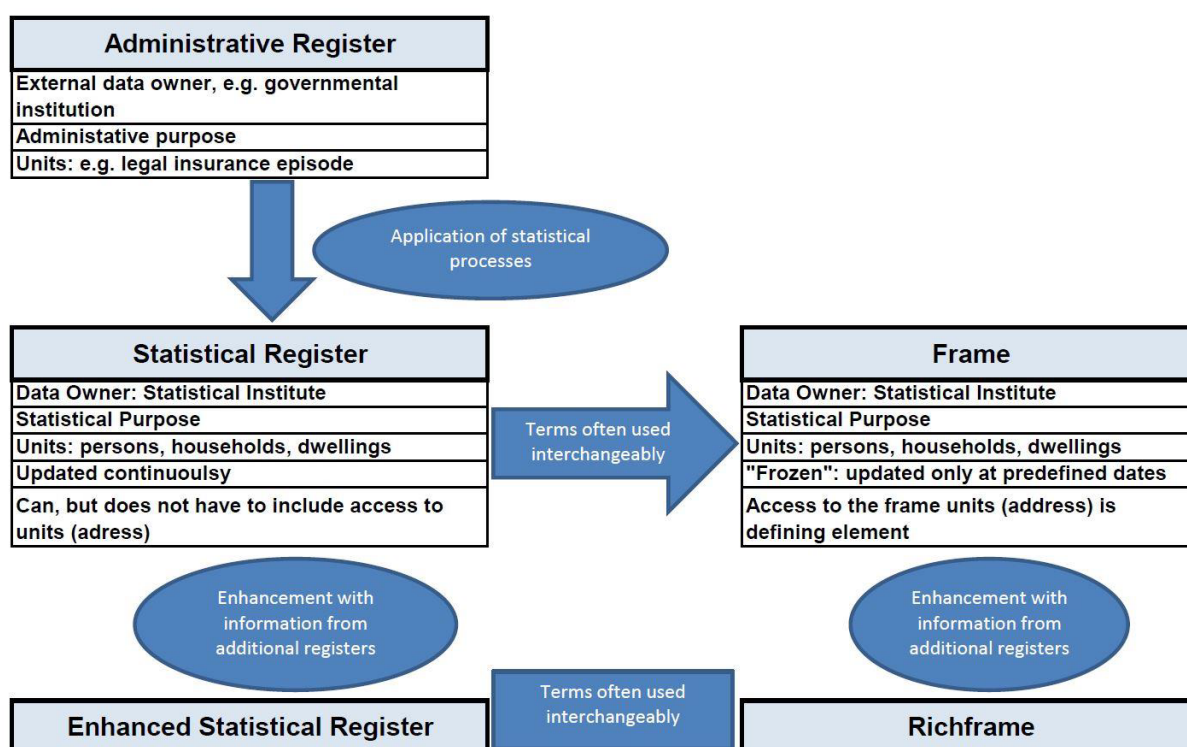


Figura 1.1: Relación y diferencias entre registros administrativos, registros estadísticos y marcos muestrales. Tomado de [Burg y col. 2019](#).

Mientras un registro administrativo es mantenido por un propietario de los datos externo a la oficina estadística para un propósito administrativo concreto (p.ej. cotizaciones

<sup>11</sup><https://stats.oecd.org/glossary/detail.asp?ID=3003>.

a la Seguridad Social), un registro estadístico está creado para propósitos estadísticos, normalmente por oficinas de estadística. Un registro estadístico se crea usualmente procesando datos de registros administrativos y/u otras fuentes de datos administrativos<sup>12</sup>. Dos de las diferencias más importantes entre los registros administrativos y estadísticos están relacionadas con la propiedad de los datos y con la unidad de análisis de interés: en el caso de los registros estadísticos, el propietario es la oficina estadística y la unidad del registro es la unidad de análisis (p.ej. una persona, un hogar, un establecimiento industrial... y no un formulario de alta, baja o modificación para recibir una subvención o un cambio en una cotización social). Los registros estadísticos se basan a menudo en registros administrativos, pero es preciso ejecutar muchos pasos del proceso de producción para que el registro estadístico satisfaga los requisitos para la producción estadística oficial.

## 1.5 Marco de áreas y otros marcos

Es importante la siguiente distinción:

- i. Un marco como una lista directa o identificación directa de elementos de la población objetivo.
- ii. Un marco como una lista o identificación de conjuntos (más grandes o más pequeños) de elementos de la población objetivo.

En el caso (i), se puede llevar a cabo un muestreo directo de elementos. En el caso (ii), el acceso a los elementos es más indirecto, concretamente, seleccionando conjuntos de elementos y observando todos o algunos de los elementos en estos conjuntos seleccionados. En muchas situaciones, el caso (ii) es la única opción, ya que no es posible encontrar o construir (sin un coste prohibitivo) una lista de elementos. El número total de elementos en la población a menudo es desconocido en el caso (ii). Por ejemplo, pensemos en la población de hogares en un gran área metropolitana. En muchas ciudades no existe nada parecido a un registro completo de hogares. Se debe considerar otras unidades muestrales distintos de los hogares.

Una forma es definir unidades muestrales como viviendas y seleccionar una muestra de este tipo de unidades. Con relativa facilidad podemos entonces conseguir acceso a los hogares en (un número reducido de) las viviendas seleccionadas. Una variante de la misma idea tiene lugar cuando se identifican segmentos dentro de un mapa forestal y se selecciona una muestra de fragmentos con el objetivo de observar árboles en los segmentos seleccionados.

---

<sup>12</sup>Se define una fuente de datos administrativos como *una tenencia de datos que contienen información recogida primariamente para propósitos administrativos (ni de investigación ni estadísticos)*. Este tipo de datos se recoge por departamentos ministeriales y otras organizaciones para los propósitos de registro, transacción y almacenamiento, normalmente durante la administración de un servicio. Incluye registros administrativos (con identificador único) pero también posiblemente otros datos administrativos sin identificador único.

El concepto de marco de áreas se define, por tanto, como [Särndal, Swensson y Wretman 1992](#):

### Definición 2

Se define un *marco de áreas* como un marco geográfico que consiste en unidades de áreas; cada elemento de la población objetivo pertenece a una unidad muestral de área y puede ser identificado tras una inspección de esta unidad de área. La unidad de área puede variar en tamaño y en el número de elementos que contiene.

El *muestreo de áreas* implica muestrear a partir de un marco de áreas, como un mapa de una ciudad, un mapa forestal o una fotografía aérea. Los conjuntos de elementos seleccionados con la ayuda de un marco de áreas a menudo se llaman *conglomerados*<sup>13</sup>. En un segundo paso de selección los conglomerados seleccionados pueden ser submuestreados. Se puede definir y muestrear una muestra de áreas incluso más pequeñas y así consecutivamente hasta que los elementos en sí mismos son finalmente muestreados en el paso final.

Los mapas, por supuesto, no siempre se usan cuando se muestrean conjuntos (conglomerados) de elementos; se puede usar en su lugar una sucesión de listas. Un marco para estudiar una población de estudiantes universitarios puede consistir en el primer paso en una lista de universidades, a continuación una lista de las facultades en las universidades seleccionadas y en el tercer y último paso se tendría acceso a los estudiantes. “Marco” aquí se refiere a un recurso con tres capas consecutivas. Las universidades son las unidades muestrales de primera etapa; las facultades, las unidades muestrales de segunda etapa y los elementos individuales (los estudiantes) son las unidades muestrales en la tercera y última etapa. En una selección que consiste de distintas etapas, cada etapa tiene su propio tipo de unidad muestral.

Una población finita está compuesta de elementos. A menudo se denominan también *unidades de análisis*, lo que recalca que son unidades que son medidas y para las cuales se graban los valores. Por ejemplo, si alguien está interesado en estimar el total poblacional de la variable ‘ingreso del hogar’, el elemento (o unidad de análisis) es naturalmente el hogar. El marco es un instrumento para acceder (de forma más o menos directa) a estos elementos. Una forma es seleccionar primero bloques de viviendas y después observar los hogares de los bloques seleccionados.

Nuestros ejemplos hasta ahora pueden haber dado la impresión de que en la práctica «elemento» es siempre algo «menor que» o al menos «igual a» una «unidad muestral». Esto no tiene por qué ser así necesariamente, como se ve en el siguiente ejemplo.

<sup>13</sup>Traducimos *cluster* como *conglomerado*.

*Ejemplo 5.* Supongamos que quiere estimarse la prevalencia en hogares de una enfermedad rara. Una posible estrategia es seleccionar los historiales médicos de enfermos en hospitales y entrevistar a padres e hijos biológicos de tales enfermos, así como todos los miembros de sus hogares. En este caso, la unidad de análisis es el hogar, pero la unidad muestral es el enfermo registrado en hospitales. Este tipo de selección de unidades es característico del llamado *muestreo en redes*<sup>14</sup> o *muestreo de multiplicidad*<sup>15</sup> (véase p.ej. [Thompson 2012](#)). ■

## 1.6 Población marco y población objetivo

Resulta necesario en este punto distinguir entre población objetivo y población marco. La *población objetivo* es el conjunto de elementos sobre los cuales queremos obtener información y para la que es necesaria estimaciones de los parámetros. La *población marco* es el conjunto de elementos que o están directamente incluidas como unidades en el marco o pueden ser identificadas mediante un concepto de marco más complejo, como un marco de selección en varias etapas.

La calidad del marco puede evaluarse considerando distintos tipos de errores ([Burg y col. 2019](#)):

- Errores de cobertura debidos a unidades muestrales faltantes, erróneamente incluidas o duplicadas.
- Errores de clasificación y de dominios de las unidades muestrales (p.ej. en el código de actividad económica principal de una empresa o en el municipio de residencia de una persona).
- Errores en la información de contacto de las unidades muestrales (p.ej. direcciones postales desactualizadas).
- Errores de alineación<sup>16</sup> en las unidades muestrales.
- Errores de unidad en las unidades muestrales compuestas (p.ej. al componer erróneamente hogares a partir de personas).

Todos estos errores son errores *ajenos al muestreo* (*non-sampling errors*) y, por tanto, afectan a la dimensión de la calidad llamada *acuracidad* (véase el Tema 18 sobre la calidad en la producción estadística oficial). La cobertura es uno de los aspectos más importantes de la calidad del marco.

En primer lugar, la *subcobertura* (*under-coverage*), en la que determinadas partes y elementos de la población objetivo no están integrados de modo sistemático y correcto, conduce a severos problemas en el uso del marco. Por ejemplo, personas viviendo en

---

<sup>14</sup>Network sampling.

<sup>15</sup>Multiplicity sampling.

<sup>16</sup>Traducimos *alignment* como alineación. Por error de alineación (*alignment error* se indica la falta de concordancia entre variables respecto de una misma unidad. Por ejemplo, entre nombres de empresas y razones sociales. Este error conlleva la aparición de más errores (como el error de unidad (*unit error*)).



el extranjero o personas sin hogar a menudo no están incluidos en numerosos procedimientos administrativos de registro de la población que alimentan el marco, que, por tanto, no las contiene. El impacto de la subcobertura en la estimación mediante estimadores lineales se hace presente, sobre todo, en el sesgo de las estimaciones. Puede hacerse una distinción entre la subcobertura por diseño y la subcobertura intrínseca. La primera se produce cuando se excluyen voluntariamente por parte del estadístico determinadas unidades muestrales (por ejemplo, porque son difíciles de localizar o su contacto tiene asociado costes muy altos). La subcobertura intrínseca se produce cuando las unidades muestrales no se encuentran en el marco por otras razones no voluntarias. Por su parte, la *sobrecobertura* (*over-coverage*) tiene lugar si existen unidades muestrales duplicadas, no existentes o fuera del ámbito de la población objetivo. Se distinguen igualmente dos tipos de sobrecobertura: listados duplicados<sup>17</sup> (*duplicate listings*) y enumeración errónea. El primer tipo corresponde a elementos de la población objetivo que están referidos al menos dos veces mediante unidades muestrales. Los duplicados afectan especialmente a la calidad de las estimaciones por su efecto a través de las variables auxiliares. Además, incrementan el coste tanto de la recogida como del procesamiento de datos. El segundo tipo hace referencia a elementos no elegibles para la población objetivo bajo análisis. Su efecto negativo, en contraposición a la subcobertura, surge en la variabilidad (varianza) de los estimadores, al reducir el tamaño muestral (al ser descartadas durante la recogida).

En segundo lugar, los errores de clasificación y de dominios equivalen a subcobertura en un dominio y sobrecobertura en otro dominio (p.ej. una empresa con un código de actividad económica principal erróneo o una persona con un municipio de residencia erróneo). Por tanto, este tipo de errores introduce información auxiliar incorrecta.

En tercer lugar, las variables de contacto en un marco desempeñan un papel fundamental para poder recoger la información en la correspondiente fase de producción. Deben estar, por tanto, monitorizadas y comprobadas periódicamente.

Por último, la relación entre las unidades muestrales es cada vez más importante. Tradicionalmente ya en las estadísticas sociales estas relaciones han sido importantes para identificar hogares y unidades compuestas de diversa naturaleza. En las encuestas económicas, aunque tradicionalmente las unidades legales han sido objeto de estudio, los grupos y *holdings* empresariales formados por varias unidades representan más recientemente un objetivo de análisis económico de creciente importancia. En este sentido, los errores de alineamiento y de unidad deben ser detectados y corregidos en la creación y mantenimiento de los marcos.

Es importante recordar que si se selecciona una muestra probabilística de un marco muestral, se puede hacer inferencia estadística válida sobre la población marco. Si la población marco difiere de la población objetivo, la inferencia sobre la población objetivo no será válida y, por tanto, el objetivo de la encuesta puede fallar. El problema es particularmente serio si el marco da acceso sólo a parte de la población objetivo. Este es

---

<sup>17</sup>En español también se usa el término *unidades repetidas*.

uno de los retos más notables para el uso de las nuevas fuentes de datos digitales en la producción estadística oficial, pues los datos se generan antes incluso de identificar las necesidades de información y objetivos de la estadística concreta.

Trabajar con un marco de muestreo perfecto no siempre es posible en la práctica. Se pueden tolerar imperfecciones leves, ya que puede no ser posible obtener un marco perfecto sin un coste excesivo. Sin embargo, es muy importante que las imperfecciones del marco sean leves. Construir un marco de alta calidad para la población objetivo es un aspecto importante de la planificación de la encuesta y debe disponerse de recursos adecuados para esta actividad. Visto de una forma diferente, cuando se define la población objetivo, debe establecerse un objetivo realista. No tiene sentido fijar una población objetivo para la que no se puede obtener un buen marco dentro de las restricciones presupuestarias. Se pueden obtener resultados inválidos a partir de muestras obtenidas de marcos erróneos. Deben evitarse marcos baratos y fáciles de construir si sólo dan acceso parcial a la población objetivo.

## 1.7 Planificación de una encuesta

Normalmente una encuesta tiene su origen en algún problema práctico. Alguien - un miembro del parlamento, un investigador, una asociación de empresarios (necesidades externas) o incluso otro departamento de la propia oficina estadística (necesidades internas) - tiene una necesidad de disponer de determinada información. Lo importante en este momento es que el problema se explique de una forma clara y concisa. Puede que esa información ya exista y únicamente sea necesario recabarla, pero puede que no exista y que sea necesario llevar a cabo una operación estadística<sup>18</sup>. La realización de una operación estadística puede llevarse a cabo usando registros administrativos, realizando un censo o una encuesta, mediante la síntesis de otras operaciones, etc<sup>19</sup>. Si se opta por la realización de una encuesta, los estadísticos deben tener claro desde el principio cuáles son los objetivos. ¿Cuál es exactamente el problema? ¿Cuál es exactamente la información necesaria?

Por ejemplo, supongamos que la propuesta es realizar una encuesta sobre las condiciones de la vivienda de los ancianos. Esta descripción es vaga y demasiado general. Se deben proporcionar definiciones claras de los conceptos involucrados. Es necesario precisar cuál es la población objetivo, a partir de qué edad se considerará a una persona «anciana» a la hora de realizar la encuesta. ¿Se debe considerar únicamente aquellas viviendas en las que viven ancianos o también aquéllas en las que personas ancianas conviven con personas de otras edades? ¿Cuál es la definición de condiciones de la

---

<sup>18</sup>Se define como operación estadística al conjunto de actividades, incluidas las preparatorias, que partiendo de una recogida de datos individuales conducen a la obtención y/o difusión de resultados estadísticos agregados, en forma de tablas o de índices, sobre un determinado tema relativo a la realidad demográfica, social, económica, ecológica, etc. de la nación o sobre un determinado territorio de ella. Véase [https://www.ine.es/GS\\_FILES/IOE\\_Metodologia.pdf](https://www.ine.es/GS_FILES/IOE_Metodologia.pdf)

<sup>19</sup>Uno de los retos presentes en la producción estadística oficial es la incorporación de nuevas fuentes de datos digitales (Big Data), como datos de la web, datos de transacciones financieras, datos de telefonía móvil, etc.



vivienda? ¿Nos referimos a la edad de la vivienda o a alguna otra medida de calidad de la vivienda? ¿Qué periodo de tiempo se estudiará? ¿Se debería distinguir entre la población anciana rural y la urbana?

A medida que se van respondiendo a estas preguntas, los estadísticos trabajan en la reformulación del proyecto inicial hacia uno en el que figuren todas las necesidades. La formulación final de estas necesidades debe clarificar los siguientes puntos:

- i. La población finita y las subpoblaciones para las cuales se requiere la información.
- ii. Los tipos de información necesarios para esta población, es decir, las variables a medir y los parámetros a estimar.

Una vez que las definiciones operacionales <sup>20</sup> se han enunciado de forma clara, los estadísticos pueden trabajar en la especificación de un diseño de encuesta adecuado, incluyendo el diseño del marco y de la muestra, el método de recogida de datos, el procesamiento y el análisis de datos (integración, codificación, depuración, imputación, estimación, validación, control del secreto estadístico, ajuste estacional y de efecto calendario, etc.) y la difusión y presentación de resultados.

De acuerdo a [Deming 1950](#), pág. 3, *la exigencia de una declaración sencilla de lo que se desea (la especificación de la encuesta) es quizá una de las mayores contribuciones de la estadística teórica moderna*. El uso de la probabilidad para la selección de las muestras y la construcción de estimadores (muestreo probabilístico) fue un hito histórico no ya desde el punto de vista matemático, sino desde el punto de vista de la gestión de la producción, porque permitió por primera vez conectar de modo directo una medida objetiva de la acuracidad de una encuesta con el coste asociado, entendiendo por coste no solo la cuestión presupuestaria, sino restricciones como el tiempo de publicación y los recursos tanto humanos como tecnológicos y logísticos. La consideración del coste en el diseño de una operación estadística es un elemento fundamental que debe estar muy presente. Antes de comenzar los trabajos de desarrollo e implementación de metodologías, tecnologías y recursos, el estadístico responsable debe asegurarse que las necesidades del usuario están claramente identificadas y el diseño podrá dar una solución, al menos suficientemente aproximada y precisa, al problema planteado.

Algunos aspectos importantes de la *planificación de una encuesta* son ([Särndal, Swensson y Wretman 1992](#)):

- Especificación de los objetivos de la encuesta.

---

<sup>20</sup>La definición de **operacionalizar** es *proceso metodológico por el que ciertos elementos del problema de estimación (sobre todo, unidades estadísticas, variables y parámetros poblacionales) se representan mediante conceptos matemáticos o estadísticos (operacionales)*. Este proceso consiste en hacer operativo un concepto abstracto o matemático. Por ejemplo, la variable parado/a, de claro interés para el usuario, no se recoge directamente en el cuestionario. Se trata de un concepto que resulta de un proceso de operacionalización sobre variables objetivo recogidas efectivamente en el cuestionario.

- Traducción del problema en el contexto temático de interés en un problema de producción de encuestas.
- Especificación de una población objetivo, variables conocidas (variables auxiliares), variables de análisis, parámetros poblacionales a estimar.
- Construcción del marco muestral, si no existe ninguno disponible que se ajuste a las necesidades de la encuesta.
- Inventario de recursos disponibles en términos de presupuesto, personal, metodología estadística, tecnologías, logística y cualquier otro equipamiento.
- Especificaciones de los requisitos a alcanzar, por ejemplo, plazos previstos de ejecución y acuracidad de las estimaciones.
- Especificación de los métodos de recogida de datos, incluyendo el diseño y construcción del cuestionario.
- Especificación del diseño muestral, mecanismo de selección de la muestra (algoritmo muestral) y determinación del tamaño muestral.
- Especificación de los métodos de integración de datos, especialmente cuando se precisa usar varias fuentes de datos.
- Especificación de clasificaciones (preferiblemente estándares) y la codificación de variables.
- Especificación de los métodos de depuración e imputación y del tratamiento de los errores ajenos al muestreo, en general.
- Especificación de los estimadores (puntuales) y las medidas de precisión (estimadores de la varianza), teniendo en cuenta también los errores ajenos al muestreo.
- Especificación del método de control del secreto estadístico.
- Especificación del ajuste estacional y de efecto calendario.
- Formación del personal y organización del trabajo de campo.
- Distribución de los recursos entre las distintas operaciones de la encuesta.
- Distribución de los recursos de control y evaluación.
- Elaboración del plan de evaluación de la calidad (indicadores, etc.).

La planificación de la encuesta debería dar lugar a una decisión para cada operación en la encuesta. La teoría estadística nos puede llevar a importantes conclusiones sobre *algunas* de estas decisiones, en particular en relación con la selección de la muestra, elección del estimador, distintas fuentes de error y sus componentes de la varianza asociadas, métodos para evaluar la acuracidad de las estimaciones y el análisis estadístico de los datos de la encuesta.

El proceso de planificación debe tratar de predecir dificultades que pueden surgir. Deben reservarse algunos recursos e identificarse procesos de *back-up* con el fin de enfrentarse con posibles dificultades. Por ejemplo, se puede esperar de forma segura alguna falta de respuesta y esto debe ser tenido en cuenta a la hora de seleccionar la muestra con el fin de que la falta de respuesta sea lo menor posible y no afecte a la calidad de las estimaciones<sup>21</sup>. Deben identificarse procedimientos de seguimiento y recontacto con los informantes que no responden y tenidos en cuenta en el presupuesto y en la previsión de plazos. Deben identificarse procedimientos metodológicos que permitan un ajuste por falta de respuesta y otros errores ajenos al muestreo en general.

De forma ideal, la planificación de la encuesta debería dar lugar a unas especificaciones óptimas para la producción de la encuesta en conjunto. El objetivo es obtener la mejor acuracidad posible sujeto a un presupuesto fijo. En una encuesta de grandes dimensiones, sin embargo, la complejidad es tan grande que no es concebible obtener una solución óptima. Hay demasiadas decisiones interrelacionadas y demasiadas variables a tener en cuenta. El concepto de *diseño total de una encuesta*<sup>22</sup>, que se verá en la siguiente Sección 1.8 puede verse como una herramienta orientada a conseguir una optimización global de una encuesta.

En la actualidad, las tareas para la planificación de una operación estadística están detalladas en las fases 1 a 3 del estándar internacional de producción GSBPM (UNECE 2019b), que incluyen la especificación de las necesidades de información, el diseño de la operación y la construcción de las herramientas de producción necesarias.

## 1.8 Diseño total de una encuesta

El término *diseño total de una encuesta* se usa en los procesos de planificación que buscan una optimización de conjunto en una encuesta. El concepto surgió del objetivo de control de conjunto sobre todas las fuentes de error en una encuesta. Hoy día también se conoce como el paradigma del error de encuesta total<sup>23</sup>. Véase, p.ej., Groves y Lyberg 2010 y múltiples referencias allí citadas.

El diseño global de una encuesta está interesado en obtener la mejor precisión posible en las estimaciones de una encuesta a la vez que busca un equilibrio económico general entre los errores de muestreo y los errores ajenos al muestreo. Para una visión general del diseño global de una encuesta, es útil considerar una encuesta desde tres perspectivas:

1. Los requisitos.
2. Las especificaciones de la encuesta.

---

<sup>21</sup>En el caso de las operaciones estadísticas económicas en España la falta de respuesta suele ser baja debido a las sanciones económicas conforme a la legislación estadística correspondiente y al seguimiento realizado.

<sup>22</sup>*Total survey design.*

<sup>23</sup>*Total survey error.*

### 3. Las tareas de producción de la encuesta.

Por requisitos nos referimos a las necesidades de información sobre la población objetivo, normalmente generadas por algún problema relacionado con algún tema social, económico, demográfico, etc. Con estos requisitos se corresponde una encuesta conceptual que alcanzará el *objetivo ideal*, si se lleva a cabo bajo las mejores circunstancias posibles.

Las especificaciones de la encuesta son un conjunto de reglas y de operaciones, que juntos constituyen un *objetivo definido* de la encuesta. Debido a las condiciones reales, este objetivo definido puede diferir del objetivo ideal. El objetivo definido especifica los elementos principales de la encuesta, como la población, el diseño muestral, los procedimientos de medida, los estimadores y las variables auxiliares.

Normalmente existen varios diseños de encuesta que nos permiten alcanzar el objetivo definido. Los estadísticos eligen de un conjunto de diseños de encuestas operativamente viables uno que se acerca lo más posible a la realización del objetivo definido. El diseño seleccionado da lugar a varias *tareas de producción de la encuesta*. La encuesta finalmente se lleva a cabo realizando estas tareas tan cuidadosamente como sea posible.

## Bibliografía

- Brewer, K.R.W. y M. Hanif (1983). *Sampling with unequal probabilities*. Springer.
- Burg, T., A. Kowarik, M. Six, G. Brancato y D. Krapavickaitė (2019). *Quality Guidelines for Frames in Social Statistics*. ESSnet KOMUSO Quality in Multisource Statistics. URL: <https://ec.europa.eu/eurostat/cros/system/files/qgfs-v1.51.pdf>.
- Deming, W.E. (1950). *Some theory of sampling*. New York: Wiley.
- Groves, R.M. y L. Lyberg (2010). "Total survey error: past, present, and future". En: *Public Opinion Quarterly* 74, págs. 849-879.
- Lessler, J.T. y W.D. Kalsbeek (1992). *Nonsampling error in surveys*. New York: Wiley.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
- Statistics Canada (2010). *Survey methods and practices*. Ottawa: Ministry of Industry of Canada. ISBN: 978-1-100-16410-6. URL: <https://www150.statcan.gc.ca/n1/en/pub/12-587-x/12-587-x2003001-eng.pdf?st=-RZ4HER2>.
- Thompson, S.K. (2012). *Sampling*. 3rd. Wiley.
- Tillé, Y. (2006). *Sampling algorithms*. Springer.
- UNECE (2019a). *Generic Statistical Information Model v1.2*. URL: <https://statswiki.unece.org/display/gsim/>.
- (2019b). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.

## Tema 2

**Ideas básicas sobre estimación en muestreo probabilístico. Diseño muestral. Probabilidades de inclusión. La noción de estadístico. Indicadores de pertenencia a la muestra. Estimadores y sus propiedades básicas. El estimador Horvitz-Thompson (estimador  $\pi$ ) y sus propiedades. Muestreo con reemplazamiento. Efecto de diseño. Intervalos de confianza.**

Este tema está elaborado como una adaptación casi literal en español del capítulo 2 de la siguiente bibliografía:

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

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.

### 2.1 Ideas básicas sobre estimación en muestreo probabilístico

El objetivo fundamental de una encuesta por muestreo consiste en obtener información sobre *características de la población, parámetros, agregados o indicadores* desconocidos a partir de la información procedente de una *muestra* representativa que permita hacer inferencia sobre la población.

Considérese una población constituida por  $N$  elementos  $\{u_1, \dots, u_N\}$ , que se denota por  $U = \{1, \dots, N\}$ , donde el tamaño de la población  $N$  es conocido.

Una muestra es un subconjunto de elementos de la población  $U$  seleccionados de un marco muestral. Asumiremos que se trata de una muestra *probabilística*, es decir, ha sido obtenida a partir de un esquema de muestreo probabilístico y se denota por  $s$ .  $S$  representará a la muestra aleatoria y  $n_S$  al número de elementos o cardinal de  $S$ .

Supongamos que se desea obtener información sobre la variable 'ingreso del hogar'. A

esta variable de interés se la denomina *variable de estudio* o *variable objetivo*. Podríamos estar interesados en obtener información sobre el total de ingresos de los hogares de la población, esto es,

$$Y_U = \sum_{k \in U} y_k$$

o bien sobre el ingreso medio

$$\bar{y}_U = \frac{1}{N} \sum_{k \in U} y_k,$$

donde  $y_k$  es el ingreso del hogar  $k$ .

El estadístico debe elegir el *diseño muestral* que desea aplicar para estimar los parámetros poblacionales total o media, por ejemplo, puede elegir un *muestreo aleatorio simple sin reemplazamiento*, diseño que produce muestras equiprobables de tamaño de muestra fijo con todos sus elementos distintos. Asimismo, debe determinar el *procedimiento de selección de la muestra*<sup>1</sup> y elegir una fórmula (*estimador*) que le permita calcular una *estimación* para el parámetro de interés, cuya elección no es independiente, ya que, como se verá más adelante, normalmente la elección del *estimador* depende del diseño seleccionado.

Un procedimiento de selección de muestras probabilísticas es un algoritmo muestral. Existen múltiples tipos de algoritmos muestrales, que pueden clasificarse en (i) enumerativos, (ii) de martingalas, (iii) secuenciales, (iv) por extracción individual, (v) eliminatorios y (vi) de rechazo (véase Tillé 2006, para los detalles). Generalmente hablando, estos algoritmos consisten en la realización secuencial de experimentos aleatorizados que producen como resultado un elemento seleccionado en la muestra tras cada experimento o bien la inclusión o exclusión en la muestra de cada elemento del marco. En el Ejemplo 6 se muestra un ejemplo del primer caso.

*Ejemplo 6.* Consideremos el siguiente procedimiento de selección de la muestra:

1. Se selecciona un elemento de los  $N$  posibles con igual probabilidad:  $\frac{1}{N}$ .
2. Se selecciona un segundo elemento de entre los  $N - 1$  restantes con igual probabilidad:  $\frac{1}{N-1}$ .
3.  $\vdots$
- n. Se selecciona un elemento de entre los  $N - n + 1$  restantes con igual probabilidad:  $\frac{1}{N-n+1}$ .

Este algoritmo secuencial es una posible forma de realizar un *muestreo aleatorio simple sin reemplazamiento* o *muestreo aleatorio simple sin reposición*, que produce una muestra de tamaño  $n$ . Consiste en realizar  $n$  experimentos aleatorizados (en este caso extracciones) donde el resultado es la selección de un elemento en la muestra. ■

<sup>1</sup>También *algoritmo muestral* (*sampling algorithm*).

Una vez extraída la muestra, los elementos son *observados*, es decir, para cada elemento de la muestra  $k$  se *mide* el valor  $y_k$  y su valor es *grabado*. Los valores grabados son usados para calcular *estimaciones* de los parámetros de interés, el ingreso total y el ingreso medio de los hogares de la población  $U$ .

## 2.2 Diseño muestral

### Definición 3

Dado un mecanismo de selección de la muestra (algoritmo muestral), se define el concepto de *diseño de muestreo*, *diseño muestral* o, simplemente, *diseño* como una función  $p(\cdot)$  que a cada muestra  $s$  le hace corresponder la probabilidad de que dicha muestra sea seleccionada,  $p(s)$ , para todo  $s$  del conjunto de posibles muestras, denotado por  $\Omega$ .

En otras palabras, el diseño muestral  $p(\cdot)$  es la función de probabilidad de la variable aleatoria  $S$ , que toma valores en  $\Omega$ :

$$\mathbb{P}(S = s) = p(s), \text{ para todo } s \in \Omega.$$

La función  $p(\cdot)$  define una función de probabilidad sobre el espacio muestral:

- i.  $p(s) \geq 0$ , para todo  $s \in \Omega$ .
- ii.  $\sum_{s \in \Omega} p(s) = 1$ .

El diseño muestral define un conjunto de muestras posibles, donde la probabilidad de selección de cada una de ellas es estrictamente positiva,  $p(s) > 0$ . El resto de muestras tendrán probabilidad nula de selección y no están en  $\Omega$ .

*Ejemplo 7.* Bajo el mecanismo de selección de la muestra definido en el Ejemplo 6, la cardinalidad del espacio muestral es  $\binom{N}{n}$  y todas las muestras tienen probabilidad igual a  $\frac{1}{\binom{N}{n}}$ . El diseño muestral es:

$$p(s) = \frac{1}{\binom{N}{n}}, \text{ para todo } s \in \Omega.$$

Se denota como diseño del *muestreo aleatorio simple sin reemplazamiento*. ■

Se puede observar que en el diseño del *muestreo aleatorio simple sin reemplazamiento* todas las muestras obtenidas tienen el mismo número de elementos,  $n$ . Sin embargo, el tamaño de muestra producido por un determinado diseño  $p(\cdot)$  no es necesariamente el mismo para todas las muestras, como es el caso del muestreo de Bernoulli. Estos diseños se estudiarán con detalle en el Tema 3.

Por otra parte, cabe destacar que diferentes mecanismos de selección de la muestra (algoritmos muestrales) pueden ser aplicados para implementar un mismo diseño muestral.

*Ejemplo 8.* Consideremos el diseño del *muestreo aleatorio simple sin reemplazamiento* e implementemos el siguiente algoritmo de selección de la muestra:

- Se selecciona un elemento de los  $N$  posibles con igual probabilidad:  $\frac{1}{N}$ . Se reemplaza el elemento obtenido.
- Se repite el paso anterior hasta que se obtengan  $n$  elementos distintos en la muestra.

■

Los algoritmos expuestos en el Ejemplo 6 y el Ejemplo 8 son dos formas de implementar el mismo diseño.

Dos de las decisiones más importantes en el diseño de una encuesta por muestreo es (i) la elección del diseño muestral y el algoritmo muestral y (ii) la elección de un estimador con el que calcular las estimaciones del parámetro poblacional de interés. Esta combinación de diseño muestral y estimador se denomina *estrategia muestral*.

## 2.3 Probabilidades de inclusión

Dada una población de  $N$  elementos  $\{u_1, \dots, u_N\}$  y un diseño muestral  $p(\cdot)$ , con  $p(s)$  la probabilidad de seleccionar la muestra  $s$ , se define la variable aleatoria *indicador de pertenencia a la muestra* del elemento  $k$ , para representar la pertenencia de un elemento a la muestra, como:

$$I_k = I_k(S) = \begin{cases} 1, & \text{si } u_k \in S, \\ 0, & \text{si } u_k \notin S. \end{cases}$$

La probabilidad de que el elemento  $u_k$  pertenezca a la muestra se puede obtener como la suma de las probabilidades de todas las muestras que contengan al elemento  $u_k$  y se denomina *probabilidad de inclusión*. Formalmente, esto se puede expresar a través de la Definición 4.

Se denotará  $u_k \in S$  o bien  $k \in S$  para indicar que el elemento  $k$  pertenece a la muestra aleatoria.

### Definición 4

La *probabilidad de inclusión*, también denominada *probabilidad de inclusión de primer orden* del elemento  $u_k$ , se define como la probabilidad de que  $u_k$  pertenezca a la muestra. Se denota por  $\pi_k$  y se puede calcular como:

$$\pi_k = \mathbb{P}(u_k \in S) = \mathbb{P}(I_k = 1) = \sum_{s \ni u_k} p(s). \quad (2.1)$$



La probabilidad de inclusión de segundo orden de los elementos  $u_k$  y  $u_l$  se define como la probabilidad de que  $u_k$  y  $u_l$  pertenezcan a la muestra. Se denota por  $\pi_{kl}$  y se puede calcular como:

$$\pi_{kl} = \mathbb{P}(u_k, u_l \in S) = \mathbb{P}(I_k \cdot I_l = 1) = \sum_{s \ni u_k, u_l} p(s). \quad (2.2)$$

*Ejemplo 9.* Se selecciona una muestra aleatoria simple sin reemplazamiento, es decir, siguiendo el diseño de muestreo definido en el Ejemplo 7. La probabilidad de inclusión de primer orden para cualquier elemento de la población es  $\frac{n}{N}$  y la probabilidad de inclusión de segundo orden para cualesquiera  $u_k$  y  $u_l$  es  $\frac{n \cdot (n-1)}{N \cdot (N-1)}$ , con  $k \neq l$ . En efecto:

$$\begin{aligned} \pi_k &= \mathbb{P}(u_k \in S) = \mathbb{P}(u_k \text{ aparezca una vez en la muestra y las } n-1 \text{ unidades} \\ &\quad \text{restantes que forman parte de la muestra no sean } u_k) = \\ &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(N-n)!(n-1)!}}{\frac{N!}{(N-n)!n!}} = \frac{n}{N}; \quad k = 1, \dots, N. \\ \pi_{kl} &= \mathbb{P}(u_k, u_l \in S) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(N-n)!(n-2)!}}{\frac{N!}{(N-n)!n!}} = \frac{n \cdot (n-1)}{N \cdot (N-1)}; \quad k \neq l = 1, \dots, N. \end{aligned}$$

■

Un diseño muestral  $p(s)$  se denomina *diseño muestral probabilístico* si cada elemento de la población tiene una probabilidad de inclusión estrictamente positiva, esto es,  $\pi_k > 0$ ,  $\forall k = 1, \dots, N$ . Todo elemento de la población tiene oportunidad de aparecer en la muestra.

En diseños muestrales directos sobre elementos de la población donde el proceso de muestreo se realiza en una única etapa y se seleccionan directamente los elementos de la población, las probabilidades de inclusión son normalmente conocidas al principio. Sin embargo, en determinados diseños más complejos, como por ejemplo, muestreo en varias etapas, el proceso de muestreo es a menudo llevado a cabo de tal manera que no es posible calcular al principio el valor de  $\pi_k \forall k \in U$ .

Por otra parte, un diseño muestral  $p(s)$  se denomina *diseño muestral medible* si cada elemento de la población tiene probabilidad de inclusión, así como probabilidad de inclusión de segundo orden estrictamente positivas, esto es,  $\pi_k > 0$ ,  $\forall k = 1, \dots, N$  y  $\pi_{kl} > 0$ ,  $\forall k \neq l = 1, \dots, N$ .

Adviértase que si  $k = l$ , se tiene  $\pi_{kl} = \pi_{kk} = \pi_k$ , ya que  $\pi_{kk} = \mathbb{P}(I_k^2 = 1) = \mathbb{P}(I_k = 1) = \pi_k$ .

## 2.4 La noción de estadístico

El objetivo del muestreo por encuestas es realizar estimaciones de los parámetros de interés a partir de los valores observados de una muestra. Por ello, se deben construir funciones matemáticas dependientes de la muestra extraída que permitan al investigador estimar dichos parámetros.

### Definición 5

Sea  $S$  una variable aleatoria (*muestra aleatoria*) que toma valores en  $\Omega$ . Un estadístico  $Q = Q(S)$  es una función real de  $S$ . La distribución de probabilidad de un estadístico se denomina *distribución en el muestreo* de  $Q$ .

Un estadístico es una variable aleatoria que toma valores distintos en función de la muestra  $s$  extraída y no depende de un parámetro desconocido. Una vez extraída una muestra  $s$ , se podrá calcular  $Q(s)$  para todo  $s$ . La información que proporciona un estadístico es muy diversa, por ejemplo, podría ofrecer información acerca de la pertenencia del elemento  $k$  a la muestra seleccionada,  $I_k(S)$ , cuál es el valor más bajo o más alto de los elementos de la muestra para una variable  $y$ , el número de veces que aparece un elemento en la muestra o el número de elementos que contiene la muestra,  $n_S$ . El tamaño de muestra se puede escribir como

$$n_S = \sum_{k \in U} I_k(S)$$

$n_S$  es otro ejemplo de estadístico, así como el total muestral, la varianza muestral y la cuasivarianza muestral.

Para el propósito de la estimación de parámetros poblacionales por muestreo, nos interesará obtener determinados estadísticos, denominados *estimadores*, cuyos valores no varíen demasiado de una muestra a otra y se concentren en torno al valor desconocido del parámetro de interés.

La esperanza y la varianza de un estadístico  $Q = Q(S)$  se definen, respectivamente, de la siguiente forma:

$$\begin{aligned} \mathbb{E}(Q) &= \sum_{s \in \Omega} p(s) \cdot Q(s) \\ \mathbb{V}(Q) &= \mathbb{E}[(Q - \mathbb{E}(Q))^2] = \\ &= \sum_{s \in \Omega} p(s) \cdot [Q(s) - \mathbb{E}(Q)]^2 \end{aligned}$$

La covarianza entre dos estadísticos  $Q_1(S)$  y  $Q_2(S)$  se define como

$$\begin{aligned} \mathbb{C}(Q_1, Q_2) &= \mathbb{E}[(Q_1 - \mathbb{E}(Q_1)) \cdot (Q_2 - \mathbb{E}(Q_2))] = \\ &= \sum_{s \in \Omega} p(s) \cdot (Q_1(s) - \mathbb{E}(Q_1)) \cdot (Q_2(s) - \mathbb{E}(Q_2)) \end{aligned}$$

## 2.5 Indicadores de pertenencia a la muestra

Los estadísticos que estamos interesados en estudiar pueden ser expresados en función de los indicadores de pertenencia a la muestra. Consideremos por ejemplo el estadístico  $Q(S) = \sum_{k \in S} y_k$ , es decir, la suma de los valores muestrales de la característica  $y$  para los elementos de la muestra aleatoria  $S$ . El estadístico se puede expresar como

$$Q(S) = \sum_{k \in U} I_k(S) \cdot y_k.$$

A continuación se definen algunas propiedades básicas del estadístico  $I_k$ .

### Proposición 1

Dado un diseño muestral  $p(s)$  se tienen los siguientes resultados para todo  $k, l = 1, \dots, N$ :

- i.  $\mathbb{E}[I_k] = \pi_k$ ;
- ii.  $\mathbb{V}(I_k) = \pi_k \cdot (1 - \pi_k)$ ;
- iii.  $\mathbb{C}(I_k, I_l) = \pi_{kl} - \pi_k \cdot \pi_l$ ,

donde  $\mathbb{E}[I_k]$  y  $\mathbb{V}(I_k)$  representan la esperanza y la varianza de  $I_k$ , respectivamente, y  $\mathbb{C}(I_k, I_l)$ , la covarianza de  $I_k$  e  $I_l$ .

### Demostración 1

$I_k$  es una variable aleatoria con distribución Bernoulli con  $\mathbb{P}(I_k = 1) = \pi_k$ , usando (2.1). Por tanto:

- i.  $\mathbb{E}[I_k] = \pi_k$ .
- ii.  $\mathbb{V}(I_k) = \mathbb{E}[I_k^2] - [\mathbb{E}[I_k]]^2 = \pi_k \cdot (1 - \pi_k)$ .
- iii.  $\mathbb{C}(I_k, I_l) = \mathbb{E}[I_k \cdot I_l] - \mathbb{E}[I_k] \cdot \mathbb{E}[I_l] = \pi_{kl} - \pi_k \cdot \pi_l$ , usando (2.2).

### Proposición 2

Dado un diseño muestral  $p(s)$  con tamaño de muestra fijo  $n$ , se tiene:

- i.  $\sum_{k \in U} \pi_k = n$ .
- ii.  $\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = n \cdot (n - 1)$ .
- iii.  $\sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = (n - 1) \cdot \pi_k$ .

**Demostración 2**

$$\begin{aligned}
\text{i. } \sum_{k \in U} \pi_k &= \sum_{k \in U} \mathbb{E}[I_k] = \mathbb{E}\left[\sum_{k \in U} I_k\right] = \mathbb{E}[n] = n. \\
\text{ii. } \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} &= \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \mathbb{E}[I_k \cdot I_l] = \mathbb{E}\left[\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} I_k \cdot I_l\right] = \mathbb{E}[n \cdot (n-1)] = n \cdot (n-1). \\
\text{iii. } \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} &= \sum_{\substack{l \in U \\ l \neq k}} \mathbb{E}[I_k \cdot I_l] = \mathbb{E}\left[I_k \cdot \left(\sum_{l \in U} I_l - I_k\right)\right] = \mathbb{E}[I_k \cdot (n - I_k)] = \\
&= n \cdot \mathbb{E}[I_k] - \underbrace{\mathbb{E}[I_k^2]}_{\mathbb{E}[I_k]} = (n-1) \cdot \mathbb{E}[I_k] = (n-1) \cdot \pi_k.
\end{aligned}$$

*Ejemplo 10.* Si se usa el diseño del *muestreo aleatorio simple sin reemplazamiento* definido en el Ejemplo 7, es sencillo comprobar los resultados de la Proposición 2.

En efecto:

$$\begin{aligned}
\text{i. } \sum_{k \in U} \pi_k &= \sum_{k \in U} \frac{n}{N} = N \cdot \frac{n}{N} = n. \\
\text{ii. } \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} &= \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{n \cdot (n-1)}{N \cdot (N-1)} = \frac{n \cdot (n-1)}{N \cdot (N-1)} \cdot N \cdot (N-1) = n \cdot (n-1). \\
\text{iii. } \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} &= \sum_{\substack{l \in U \\ l \neq k}} \frac{n \cdot (n-1)}{N \cdot (N-1)} = (N-1) \cdot \frac{n \cdot (n-1)}{N \cdot (N-1)} = (n-1) \cdot \frac{n}{N} = (n-1) \cdot \pi_k.
\end{aligned}$$

■

## 2.6 Estimadores y sus propiedades básicas

Sea  $\theta$  un vector  $j$ -dimensional que representa los parámetros poblacionales de interés, esto es,  $\theta = (\theta_1, \dots, \theta_j)$ . Un estimador es un estadístico propuesto para producir valores (estimaciones puntuales) del parámetro poblacional  $\theta$ , que tomará distintos valores en función de la muestra elegida. Así, un estimador se describe en función de los valores muestrales  $\theta = \theta(S)$ .

Consideremos por ejemplo que se desea estimar el total poblacional de la variable  $y$ ,  $\theta(y_1, \dots, y_N) = Y_U = \sum_{k \in U} y_k$ , podemos definir el estimador  $\hat{\theta}(S) = \frac{N}{n} \sum_{k \in S} y_k$ , usado en un diseño de muestreo con probabilidades iguales y sin reemplazamiento.

**Definición 6**

Un estimador puntual del parámetro poblacional  $\theta$  es una función  $\hat{\theta}$  que a cada posible muestra  $S$  le hace corresponder una estimación  $\hat{\theta}(S)$  de  $\theta$ .

Es importante distinguir entre estimador y estimación. Una estimación es el valor  $\hat{\theta}(s)$  que se puede calcular para una muestra particular  $s$  del conjunto de muestras posibles  $\Omega$ . Por ejemplo, en el caso de un muestreo con probabilidades iguales y sin reemplazamiento, la variable aleatoria

$$\hat{\theta}(S) = \frac{N}{n} \sum_{k \in S} y_k = \frac{N}{n} \sum_{k \in U} I_k \cdot y_k$$

es un estimador del parámetro poblacional  $\theta = Y_U = \sum_{k \in U} y_k$ ; mientras que la estimación obtenida para una muestra particular seleccionada es el valor

$$\hat{\theta}(s) = \frac{N}{n} \sum_{k \in s} y_k.$$

**Comentario 1.** A partir de ahora y en los temas sucesivos se ignorará la distinción tipográfica entre  $S$ , la muestra aleatoria, y  $s$ , una muestra particular seleccionada del conjunto de muestras posibles. Por simplicidad, se usará  $s$ .

■

Al conjunto de posibles valores que puede tomar el estimador  $\hat{\theta}$  junto a la probabilidad de que  $\hat{\theta}$  tome dichos valores bajo un diseño muestral  $p(s)$  se conoce como la distribución del estimador  $\hat{\theta}$  en el muestreo.

**Definición 7**

Sea  $\hat{\theta}$  un estimador para  $\theta$  y  $C$  el conjunto de todos los posibles valores que produce el estimador. Para cada valor  $c \in C$ , la probabilidad de que el estimador tome dicho valor viene dado por

$$P_C = \mathbb{P}(\hat{\theta} = c) = \sum_{s \in \Omega_C} p(s)$$

donde  $\Omega_C$  es el conjunto de muestras  $s$  para las que  $\hat{\theta}(s) = c$ .

La distribución del estimador en el muestreo es el par  $\{C, P_C\}$ .

A continuación se definen algunas propiedades importantes de los estimadores, el concepto de *insesgadez*, *error cuadrático medio*, *error de muestreo* o *error estándar* y *error relativo* o *coeficiente de variación* del estimador.

**Definición 8**

Un estimador  $\hat{\theta}$  es insesgado para el parámetro  $\theta$  si su sesgo es 0, esto es:

$$\mathbb{B}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = 0,$$

donde la esperanza del estimador se define como

$$\mathbb{E}[\hat{\theta}] = \sum_{s \in \Omega} p(s) \cdot \hat{\theta}(s).$$

**Definición 9**

El error cuadrático medio<sup>a</sup> del estimador  $\hat{\theta}$  se define como

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[ \left( \hat{\theta} - \theta \right)^2 \right] = \sum_{s \in \Omega} p(s) \cdot \left( \hat{\theta}(s) - \theta \right)^2.$$

<sup>a</sup>Mean square error.

**Definición 10**

La varianza del estimador se define como

$$\mathbb{V}(\hat{\theta}) = \sum_{s \in \Omega} p(s) \cdot \left[ \hat{\theta}(s) - \mathbb{E}[\hat{\theta}] \right]^2.$$

Se puede demostrar el siguiente resultado:

**Proposición 3**

$$\text{MSE}(\hat{\theta}) = \mathbb{V}(\hat{\theta}) + \left[ \mathbb{B}(\hat{\theta}) \right]^2.$$

**Demostración 3**

En efecto:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \mathbb{E} \left[ \left( \hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta \right)^2 \right] = \\ &= \mathbb{E} \left[ \left( \hat{\theta} - \mathbb{E}[\hat{\theta}] \right)^2 \right] + \left[ \mathbb{E}[\hat{\theta}] - \theta \right]^2 = \\ &= \mathbb{V}(\hat{\theta}) + \left[ \mathbb{B}(\hat{\theta}) \right]^2 \end{aligned}$$

*Ejemplo 11.* Consideremos el diseño de muestreo con probabilidades iguales y sin reemplazamiento para el cual se había propuesto el estimador

$$\hat{\theta} = \frac{N}{n} \sum_{k \in s} y_k.$$

Este estimador es insesgado para  $\theta = \sum_{k \in U} y_k$ :

$$\mathbb{E}[\hat{\theta}] = \mathbb{E} \left[ \frac{N}{n} \sum_{k \in U} I_k \cdot y_k \right] = \frac{N}{n} \sum_{k \in U} \mathbb{E}[I_k] \cdot y_k = \frac{N}{n} \sum_{k \in U} \pi_k \cdot y_k \underbrace{=}_{\pi_k = \frac{n}{N}} \sum_{k \in U} y_k = \theta.$$

■

El sesgo y el error cuadrático medio son medidas importantes de la calidad de un estimador. En general, entre varios estimadores posibles para estimar el parámetro  $\theta$ , se escogerá aquel cuya distribución en el muestreo se concentre más estrechamente alrededor del verdadero parámetro  $\theta$ . Podríamos usar entonces como criterio de selección aquel estimador que tenga el menor error cuadrático medio, ya que hay razón de peso para creer que si el error cuadrático medio del estimador es bajo, la muestra extraída producirá una estimación cercana al valor verdadero. El error cuadrático medio se puede reducir a través de la varianza del estimador o bien reduciendo el sesgo. Normalmente el investigador se encargará de tomar un estimador que sea insesgado o aproximadamente insesgado y escogerá aquel que tenga menor varianza.

Como ya se comentó en el Tema 1, en el proceso de estimación de un parámetro de la población a través de muestreo probabilístico se origina un error de estimación. El error de estimación se define como la desviación de la estimación respecto al verdadero valor del parámetro desconocido que se desea estimar. El error de estimación es debido a dos tipos de errores: el *error de muestreo*, originado al observar los parámetros de interés sobre un subconjunto de la población en lugar de sobre la población entera, y otro tipo de errores, los denominados errores *ajenos al muestreo*.

El *error de muestreo* del estimador se debe exclusivamente al cálculo de la estimación a partir de los datos de un subconjunto de la población (*muestra*).

### Definición 11

El *error de muestreo* o *error estándar* del estimador  $\theta$  se define como la raíz cuadrada de la varianza del estimador, denominado  $\sigma(\hat{\theta})$ :

$$\sigma(\hat{\theta}) = \left[ \mathbb{V}(\hat{\theta}) \right]^{1/2}.$$

Por su parte, los errores *ajenos al muestreo* son el resto de errores que se pueden producir a lo largo de la investigación estadística: deficiencias en el marco muestral, falta de respuesta, errores de medida y errores de procesamiento de la información. Los errores *ajenos al muestreo* pueden producirse en cualquiera de las siguientes fases de la encuesta:

- i. Fase de selección de la muestra. Esta fase consiste en obtener, siguiendo el diseño muestral seleccionado para la encuesta, una muestra de elementos a partir de la utilización de un marco muestral adecuado. Los errores de estimación asociados a esta fase son el error de muestreo y las posibles deficiencias existentes en el marco, de las cuales destaca la *subcobertura* que produce serios problemas, ya que existen elementos de la población objetivo que no están presentes en el marco, por lo que no pueden ser seleccionados, siendo nula su *probabilidad de inclusión* en la muestra (véase el Tema 1 sobre la calidad del marco).
- ii. Fase de recogida de datos. Esta fase consiste en implementar el plan de recogida de datos para la muestra seleccionada. Se pueden producir errores debido a la falta de respuesta y errores de medida. La falta de respuesta se da, por ejemplo, cuando existe una negativa o incapacidad para responder por parte del informante, o bien el informante no se encuentra en su domicilio en el momento de la entrevista. Las principales fuentes de los errores de medida son el entrevistador (defectos en la labor de los entrevistadores por falta de formación, interpretación o grabación incorrecta de las respuestas dadas por el informante), el informante (respuestas incorrectas de forma intencionada o no, interpretación incorrecta de las preguntas del cuestionario), el cuestionario (diseño incorrecto o inadecuado) y el modo de la entrevista (véase p.ej. [Groves 1989](#)).
- iii. Fase de procesamiento de datos. En esta fase se procesa y prepara la información recogida para la fase de estimación y análisis (fase iv). Incluye la codificación de los datos, es decir, la transcripción del cuestionario a un medio adecuado para la fase iv, el proceso de depuración de los datos, mediante la implementación de técnicas de detección y corrección de errores y *outliers*, la imputación de datos faltantes y el recontacto con los informantes para clarificar cualquier tipo de información en caso de ser necesario y no suponer un coste demasiado elevado. Los errores que pueden surgir de esta fase incluyen por ejemplo errores de transcripción, codificación y errores en los valores imputados.

En la fase de estimación y análisis, además de realizar los cálculos de las estimaciones de los parámetros poblacionales, se obtienen medidas de precisión de estas estimaciones, como por ejemplo la estimación de la varianza del estimador o del error de muestreo,  $\hat{V}(\hat{\theta})$  o  $\hat{\sigma}(\hat{\theta})$ , respectivamente.

Otro indicador de precisión que a menudo suele obtenerse en las encuestas es una estimación del error relativo o *coeficiente de variación* del estimador.



**Definición 12**

El *error relativo* o *coeficiente de variación* del estimador se define como el cociente entre el error de muestreo y el valor esperado del estimador, esto es:

$$CV(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{\mathbb{E}[\hat{\theta}]}.$$

El estimador comúnmente usado para estimar el coeficiente de variación cuando el estimador  $\theta$  es insesgado o aproximadamente insesgado es:

$$cve(\hat{\theta}) = \frac{\hat{\sigma}(\hat{\theta})}{\hat{\theta}}.$$

**2.7 El estimador Horvitz-Thompson (estimador  $\pi$ ) y sus propiedades**

Sea el total poblacional de una variable  $y$  el parámetro de interés

$$Y_U = \sum_{k \in U} y_k.$$

**Definición 13**

Se denomina *estimador  $\pi$*  o estimador de Horvitz-Thompson de  $Y_U$  a:

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} I_k \cdot \frac{y_k}{\pi_k}.$$

El coeficiente  $\frac{1}{\pi_k}$  que multiplica a cada elemento se denomina *factor de expansión* (también denominado *peso de muestreo* o *factor de elevación*) e incrementa la importancia de los elementos en la muestra, de tal forma que *podría decirse* que el elemento  $k$ -ésimo, cuando está presente en la muestra, representa no solo a dicho elemento sino a  $\frac{1}{\pi_k}$  elementos de la población. Así, el factor de expansión *podría interpretarse* como el número de elementos en la población que representa cada unidad de la muestra.

*Ejemplo 12.* El estimador  $\hat{Y}_U^{\text{HT}}$  para el diseño *muestral aleatorio simple sin reemplazamiento* es

$$\hat{Y}_U^{\text{HT}} = \frac{N}{n} \sum_{k \in s} y_k.$$

■

**Teorema 4**

El estimador HT

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

es insesgado para el total  $Y_U = \sum_{k \in U} y_k$  y la expresión de la varianza del estimador es

$$\mathbb{V}(\hat{Y}_U^{\text{HT}}) = \sum_{k \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l) \quad (2.3)$$

Si  $\pi_{kl} > 0$  para todo  $k, l \in U$ , un estimador insesgado de  $\mathbb{V}(\hat{Y}_U^{\text{HT}})$  viene dado por la siguiente expresión debida a [Horvitz y Thompson 1952](#):

$$\hat{\mathbb{V}}(\hat{Y}_U^{\text{HT}}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad (2.4)$$

**Demostración 4**

El estimador  $\hat{Y}_U^{\text{HT}}$  es insesgado. En efecto:

$$\mathbb{E}[\hat{Y}_U^{\text{HT}}] = \mathbb{E}\left[\sum_{k \in U} \frac{y_k}{\pi_k} \cdot I_k\right] = \sum_{k \in U} \frac{y_k}{\pi_k} \cdot \mathbb{E}[I_k] = \sum_{k \in U} \frac{y_k}{\pi_k} \cdot \pi_k = \sum_{k \in U} y_k = Y_U.$$

La expresión de la varianza de  $\hat{Y}_U^{\text{HT}}$  es:

$$\begin{aligned} \mathbb{V}(\hat{Y}_U^{\text{HT}}) &= \mathbb{V}\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) = \mathbb{V}\left(\sum_{k \in U} \frac{y_k}{\pi_k} \cdot I_k\right) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \cdot \mathbb{V}(I_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \mathbb{C}\left(\frac{y_k}{\pi_k} \cdot I_k, \frac{y_l}{\pi_l} \cdot I_l\right) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \cdot \mathbb{V}(I_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \cdot \mathbb{C}(I_k, I_l) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \cdot \pi_k(1 - \pi_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l) = \\ &= \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \end{aligned}$$

Por último se demuestra la insesgades del estimador de la varianza,  $\hat{\mathbb{V}}(\hat{Y}_U^{\text{HT}})$ . Usando el indicador de pertenencia de un elemento a la muestra y siempre que  $\pi_{kl} > 0$

para todo  $k, l \in U$ , se puede escribir

$$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\text{HT}}\right) = \sum_{j \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot I_k \cdot I_l$$

cuya esperanza es

$$\begin{aligned} \mathbb{E}\left[\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\text{HT}}\right)\right] &= \mathbb{E}\left[\sum_{j \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot I_k \cdot I_l\right] = \\ &= \sum_{j \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \mathbb{E}[I_k \cdot I_l] = \\ &= \sum_{j \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \pi_{kl} = \mathbb{V}\left(\widehat{Y}_U^{\text{HT}}\right). \end{aligned}$$

### Corolario 5

Dado un diseño muestral  $p(s)$  con un tamaño de muestra fijo  $n$ , entonces la varianza del estimador  $\pi$  puede ser también escrita de la siguiente forma:

$$\mathbb{V}(\widehat{Y}_U^{\text{HT}}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \quad (2.5)$$

Si  $\pi_{kl} > 0$  para todo  $k, l \in U$ , un estimador insesgado de  $\mathbb{V}(\widehat{Y}_U^{\text{HT}})$  viene dado por la siguiente expresión debida a [Yates y Grundy 1953](#) y [Sen 1953](#):

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \quad (2.6)$$

### Demostración 5

Se puede demostrar que las expresiones (2.3) y (2.5) son equivalentes cuando el tamaño del diseño es fijo. En efecto:

$$\begin{aligned} \mathbb{V}(\widehat{Y}_U^{\text{HT}}) &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \left[ \left(\frac{y_k}{\pi_k}\right)^2 - 2 \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} + \left(\frac{y_l}{\pi_l}\right)^2 \right] = \\ &= \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} - \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \left(\frac{y_k}{\pi_k}\right)^2 \end{aligned}$$

donde  $\sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \left(\frac{y_k}{\pi_k}\right)^2 = \sum_{k \in U} \left(\frac{y_k}{\pi_k}\right)^2 \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) = 0$ , ya que:

$$\begin{aligned} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) &= \sum_{l \in U} \pi_{kl} - \pi_k \sum_{l \in U} \pi_l = \sum_{l \in U} \mathbb{E}[I_k \cdot I_l] - \pi_k \sum_{l \in U} \mathbb{E}[I_l] = \\ &= \mathbb{E} \left[ \sum_{l \in U} I_k \cdot I_l \right] - \pi_k \cdot \mathbb{E} \left[ \sum_{l \in U} I_l \right] = \mathbb{E}[n \cdot I_k] - \pi_k \cdot \mathbb{E}[n] = \\ &= n \cdot \pi_k - n \cdot \pi_k = 0 \end{aligned}$$

Se obtiene así la expresión (2.3).

En cuanto a la insesgader del estimador de la varianza, si  $\pi_{kl} > 0$  para todo  $k, l \in U$ , entonces  $\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}})$  se puede escribir como

$$\begin{aligned} \widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}}) &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} I_l \cdot I_l \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_k}{\pi_{kl}} \cdot \left(\frac{y_k}{\pi_l} - \frac{y_l}{\pi_l}\right)^2 \\ \mathbb{E}[\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}})] &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \mathbb{E}[I_l \cdot I_l] \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_k}{\pi_{kl}} \cdot \left(\frac{y_k}{\pi_l} - \frac{y_l}{\pi_l}\right)^2 = \\ &= -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} \pi_{kl} \cdot \frac{\pi_{kl} - \pi_k \cdot \pi_k}{\pi_{kl}} \cdot \left(\frac{y_k}{\pi_l} - \frac{y_l}{\pi_l}\right)^2 = \mathbb{V}(\widehat{Y}_U^{\text{HT}}) \end{aligned}$$

Como se ha comentado, las dos expresiones de la varianza del estimador (2.3) y (2.5) son idénticas cuando el diseño produce muestras de tamaño fijo. Sin embargo, los dos estimadores insesgados de las varianzas (2.4) y (2.6) no son necesariamente iguales.

Puede comprobarse que para el diseño del *muestreo aleatorio simple sin reemplazamiento*, que es un diseño que produce muestras de tamaño fijo, las dos expresiones de la varianza del estimador dado en el Ejemplo 12 y del estimador de la varianza producen los mismos resultados.

## 2.8 Muestreo con reemplazamiento

Los esquemas de muestreo con reemplazamiento permiten seleccionar elementos ya extraídos previamente, mientras que en el esquema sin reemplazamiento si un elemento ha sido extraído no podrá ser seleccionado nuevamente.

Supongamos que la probabilidad de que un elemento de la población sea seleccionado es  $p_k > 0$ ,  $\forall k = 1, \dots, N$ , con  $\sum_{k \in U} p_k = 1$  y se realizan  $n$  extracciones con reemplazamiento. El esquema general de un proceso de muestreo con reemplazamiento consiste en seleccionar el primer elemento de la muestra, por ejemplo  $k_1$ , y devolverlo nuevamente al conjunto de posibles elementos a seleccionar. A continuación se extrae el segundo

elemento de la muestra,  $k_2$ , y se vuelve a reponer, y así sucesivamente hasta realizar  $n$  extracciones.

Se puede observar que las  $n$  extracciones son independientes y las probabilidades de seleccionar una unidad son las mismas en cada extracción. Podemos escribir la muestra resultante como una *muestra ordenada* de  $n$  elementos como sigue:

$$os = (k_1, \dots, k_n).$$

La *muestra ordenada* contiene información sobre el orden de extracción y el número de veces que cada elemento aparece en la muestra. Sin embargo, la información sobre el orden de extracción no interesa para nuestro propósito, ya que muestras con los mismos elementos pero en distinto orden las consideraremos la misma. Así, para un diseño de tamaño fijo  $n$ , el número posible de muestras distintas es una combinación con repetición de  $N$  elementos tomados de  $n$  en  $n$ .

Se define la variable aleatoria  $m_k$  como el número de veces que aparece el elemento  $k$  en la muestra con  $k = 1, \dots, N$ , cuya distribución es binomial de parámetros  $n$  y  $p_k$ . La probabilidad de seleccionar una muestra viene dada por el modelo multinomial:

$$\mathbb{P}(e_1 = m_1, \dots, e_N = m_N) = \frac{n!}{m_1! \cdot \dots \cdot m_N!} \cdot (p_1)^{m_1} \cdot \dots \cdot (p_N)^{m_N},$$

donde  $\sum_{k \in U} m_k = n$ .

La probabilidad de inclusión del elemento  $k$ ,  $\forall k \in U$  se puede calcular de la siguiente forma:

$$\begin{aligned} \pi_k &= \mathbb{P}(u_k \in s) = \mathbb{P}(m_k \neq 0) = 1 - \mathbb{P}(m_k = 0) = \\ &= 1 - \binom{n}{0} \cdot p_k^0 \cdot (1 - p_k)^n = 1 - (1 - p_k)^n \end{aligned}$$

*Ejemplo 13.* Consideremos el esquema de muestreo consistente en la extracción independiente de  $n$  elementos de forma que cada elemento tiene la misma probabilidad de ser seleccionado, esto es,  $p_k = \frac{1}{N}$ ,  $\forall k = 1, \dots, N$ . Cada vez que se extrae un elemento se repone en la población de tal forma que todos los elementos pueden ser seleccionados en cada extracción. Este mecanismo de selección de la muestra es una posible implementación del diseño denominado *muestreo aleatorio simple con reemplazamiento* o *muestreo aleatorio simple con reposición*.

Bajo este diseño, las probabilidades de inclusión de primer y segundo orden pueden obtenerse de forma sencilla, respectivamente:

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^n, \quad k = 1, \dots, N;$$

$$\pi_{kl} = 1 - 2 \cdot \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n, \quad k \neq l = 1, \dots, N.$$

**Definición 14**

Sea  $Y_U$  el total poblacional de la variable  $y$  el parámetro de interés. El estimador de  $Y_U$  debido a [Hansen y Hurwitz 1943](#) es

$$\hat{Y}_U^{\text{HH}} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}}$$

y lo denominaremos el *estimador HH*.

**Teorema 6**

El estimador HH

$$\hat{Y}_U^{\text{HH}} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}}$$

es un estimador insesgado para el total  $Y_U = \sum_{k \in U} y_k$  y la expresión de la varianza del estimador es

$$\mathbb{V}(\hat{Y}_U^{\text{HH}}) = \frac{1}{n} \sum_{k \in U} \left( \frac{y_k}{p_k} - Y_U \right)^2 \cdot p_k. \quad (2.7)$$

Un estimador insesgado de  $\mathbb{V}(\hat{Y}_U^{\text{HH}})$  viene dado por

$$\hat{\mathbb{V}}(\hat{Y}_U^{\text{HH}}) = \frac{1}{n \cdot (n-1)} \sum_{i=1}^n \left( \frac{y_{k_i}}{p_{k_i}} - \hat{Y}_U^{\text{HH}} \right)^2. \quad (2.8)$$

**Demostración 6**

Se define  $Z_i = \frac{y_{k_i}}{p_{k_i}}$  como una variable aleatoria que toma los valores  $Z_i = \frac{y_k}{p_k}$  si  $k_i = k$  para todo  $k = 1, \dots, N$ , es decir, si el elemento  $k$  se ha seleccionado en la extracción  $i$ -ésima. Así pues, se dispone de  $n$  variables aleatorias cuya distribución de probabilidad viene dada por

$$\mathbb{P} \left( Z_i = \frac{y_k}{p_k} \right) = \mathbb{P}(k_i = k) = p_k, \quad k \in U.$$

El estimador HH puede escribirse en función de  $Z_i$ :

$$\hat{Y}_U^{\text{HH}} = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{Z}$$

Las variables aleatorias  $Z_1, \dots, Z_n$  son independientes e idénticamente distribuidas, ya que se realizan extracciones independientes con las mismas probabilidades de selección  $(p_1, \dots, p_N)$  en cada extracción.

$$\begin{aligned}\mathbb{E}[Z_i] &= \sum_{k \in U} \frac{y_k}{p_k} \cdot \mathbb{P}\left(Z_i = \frac{y_k}{p_k}\right) = \sum_{k \in U} \frac{y_k}{p_k} \cdot p_k = Y_U \\ \mathbb{V}(Z_i) &= \mathbb{E}[(Z_i - \mathbb{E}[Z_i])^2] = \mathbb{E}[(Z_i - Y_U)^2] = \sum_{k \in U} \left(\frac{y_k}{p_k} - Y_U\right)^2 \cdot p_k\end{aligned}$$

Dado que  $\hat{Y}_U^{\text{HH}}$  es la media aritmética de  $n$  variables aleatorias independientes e idénticamente distribuidas, se tiene:

$$\begin{aligned}\mathbb{E}[\hat{Y}_U^{\text{HH}}] &= \frac{1}{n} \sum_{i=1}^n Y_U = Y_U \\ \mathbb{V}(\hat{Y}_U^{\text{HH}}) &= \frac{1}{n} \sum_{k \in U} \left(\frac{y_k}{p_k} - Y_U\right)^2 \cdot p_k\end{aligned}$$

En cuanto a la insesgadez de  $\hat{\mathbb{V}}(\hat{Y}_U^{\text{HH}})$ , dado que  $Z_1, \dots, Z_n$  son variables aleatorias independientes e idénticamente distribuidas, se tiene que la cuasivarianza muestral

$$\frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_{k_i}}{p_{k_i}} - \hat{Y}_U^{\text{HH}}\right)^2$$

es un estimador insesgado de la varianza de  $Z_i$ , quedando demostrada así la insesgadez de  $\hat{\mathbb{V}}(\hat{Y}_U^{\text{HH}})$ .

*Ejemplo 14.* Supongamos que se toma  $p_k$  proporcional a  $y_k$ , esto es,

$$p_k = c \cdot y_k, \quad k = 1, \dots, N$$

para alguna constante  $c$ . Dado que  $\sum_{k \in U} p_k = 1$ , entonces  $c = \frac{1}{\sum_{k \in U} y_k}$ .

El estimador HH es el total  $Y_U$  y se tiene  $\mathbb{V}(\hat{Y}_U^{\text{HH}}) = 0$ . ■

En la práctica no es posible seleccionar  $p_k$  proporcional a  $y_k$  puesto que estos valores normalmente no son conocidos. Sin embargo, en ocasiones son conocidos los valores  $x_1, \dots, x_k$  para una variable  $x$  cuyo ratio  $\frac{y_k}{x_k}$  es aproximadamente constante para todos los elementos, entonces puede definirse  $p_k = \frac{x_k}{\sum_{k \in U} x_k}$ , obteniendo de esta forma un procedimiento que produce una varianza pequeña. Este procedimiento se conoce como *muestreo con probabilidades proporcionales a los tamaños*.

## 2.9 Efecto de diseño

### Definición 15

Consideremos dos diseños muestrales  $p_1(s)$  y  $p_2(s)$  para estimar un parámetro poblacional  $\theta$ . Sean  $\hat{\theta}_1$  y  $\hat{\theta}_2$  los estimadores considerados, asociados a los diseños  $p_1(\cdot)$  y  $p_2(\cdot)$ , respectivamente. Se define el *efecto de diseño* como el cociente

$$\text{deff}(p_1, \hat{\theta}_1) = \frac{\mathbb{V}_{p_1}(\hat{\theta}_1)}{\mathbb{V}_{p_2}(\hat{\theta}_2)},$$

donde  $\mathbb{V}_{p_1}(\hat{\theta}_1)$  y  $\mathbb{V}_{p_2}(\hat{\theta}_2)$  son las varianzas de los estimadores  $\hat{\theta}_1$  y  $\hat{\theta}_2$ , respectivamente.

El *efecto de diseño* permite comparar la precisión obtenida por la *estrategia* compuesta por el diseño muestral  $p_1$  y el estimador  $\hat{\theta}_1$ , denotada por  $(p_1, \hat{\theta}_1)$ , en relación a otra *estrategia* formada por la selección del diseño muestral  $p_2$  y el estimador  $\hat{\theta}_2$ ,  $(p_2, \hat{\theta}_2)$ .

Si  $\text{deff}(p_1, \hat{\theta}_1) > 1$  se pierde precisión por no usar la *estrategia*  $(p_2, \hat{\theta}_2)$ . Si  $\text{deff}(p_1, \hat{\theta}_1) < 1$  se gana precisión comparado con usar  $(p_2, \hat{\theta}_2)$ .

El objetivo general es encontrar la mejor *estrategia*, es decir, aquella combinación de diseño muestral y estimador que permita estimar el parámetro poblacional desconocido de la forma más acurada posible.

*Ejemplo 15.* Consideremos el diseño muestral del *muestreo aleatorio simple sin reemplazamiento*,  $p_1(s)$ , y *con reemplazamiento*,  $p_2(s)$ . Para el primer diseño muestral se selecciona el estimador HT y para el segundo el estimador HH. En el Tema 3 se puede consultar la demostración de que las expresiones de la varianza de  $\hat{Y}_U^{\text{HT}}$  y  $\hat{Y}_U^{\text{HH}}$  son, respectivamente:

$$\begin{aligned}\mathbb{V}_{p_1}(\hat{Y}_U^{\text{HT}}) &= N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{S_{yU}^2}{n}, \\ \mathbb{V}_{p_2}(\hat{Y}_U^{\text{HH}}) &= N \cdot (N-1) \cdot \frac{S_{yU}^2}{n},\end{aligned}$$

donde  $S_{yU}^2$  es la cuasivarianza poblacional de  $y$ .

El efecto de diseño de  $p_2$  comparado con  $p_1$  es:

$$\frac{\mathbb{V}_{p_2}(\hat{Y}_U^{\text{HH}})}{\mathbb{V}_{p_1}(\hat{Y}_U^{\text{HT}})} = \frac{N \cdot (N-1) \cdot \frac{S_{yU}^2}{n}}{N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{S_{yU}^2}{n}} = \frac{N-1}{N \cdot \left(1 - \frac{n}{N}\right)} = \frac{N-1}{N-n} \geq 1, \text{ para todo } n \geq 1.$$

Como cabía esperar, el muestreo sin reemplazamiento es más eficiente (salvo si el tamaño de la muestra es  $n = 1$ ). ■



## 2.10 Intervalos de confianza

Un intervalo de confianza para el parámetro  $\theta$  es un intervalo aleatorio de la forma

$$IC_{\theta}(s) = [L(s), U(s)]$$

donde  $L$  y  $U$  son dos estadísticos tales que  $L(s) \leq U(s)$  para cada  $s$ .

Se desea que la probabilidad de que el intervalo  $IC_{\theta}(s)$  contenga el parámetro  $\theta$  sea cercana a la unidad. Esta probabilidad se denomina *nivel de confianza* y se representa como

$$\mathbb{P}[IC_{\theta}(s) \ni \theta] = 1 - \alpha,$$

donde  $\alpha$  es la probabilidad acumulada de aquellas muestras para las cuales el intervalo de confianza no incluye al valor verdadero  $\theta$ , esto es:

$$\alpha = \sum_{s \in \Omega_0} p(s),$$

donde  $p(\cdot)$  representa el diseño muestral y  $\Omega_0$  el conjunto de muestras para las cuales el intervalo de confianza obtenido no incluye a  $\theta$ .

Una vez extraída una muestra,  $s$ , el intervalo de confianza para el parámetro  $\theta$  se puede calcular y viene dado por

$$IC_{\theta}(s) = [L(s), U(s)].$$

*Ejemplo 16.* Sea  $\hat{y}$  un estimador para el total poblacional  $Y_U$  con distribución normal cuya esperanza es  $Y_U$  y su varianza es  $\mathbb{V}(\hat{y})$ , conocida. Un intervalo de confianza para  $Y_U$  viene dado por:

$$IC_y(s) = \left[ \hat{y} \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(\hat{y})} \right]$$

donde  $z_{\alpha/2}$  es el valor que verifica  $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$  donde  $Z$  es una variable aleatoria con distribución  $N(0, 1)$ . ■

*Ejemplo 17.* Sea  $\hat{y}$  un estimador para el total  $y$  con distribución normal cuya esperanza es  $y$  y su varianza es  $\mathbb{V}(\hat{y})$ , desconocida. Un intervalo de confianza para  $y$  viene dado por:

$$IC_y(s) = \left[ \hat{y} \pm t_{n-1; \alpha/2} \cdot \sqrt{\hat{\mathbb{V}}(\hat{y})} \right]$$

donde  $t_{n-1; \alpha/2}$  es el valor que verifica  $\mathbb{P}(T > t_{n-1; \alpha/2}) = \alpha/2$  considerando  $T$  una variable aleatoria con distribución  $t$  de Student con  $n - 1$  grados de libertad. ■

En los ejemplos anteriores se ha podido obtener un intervalo de confianza con nivel de confianza  $1 - \alpha$ . Sin embargo, normalmente resulta complicado obtener intervalos de confianza con nivel de confianza exacto  $1 - \alpha$ . A menudo se suele calcular el siguiente intervalo de confianza para el parámetro  $\theta$ :

$$IC_{\theta}(s) = \left[ \hat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\hat{\mathbb{V}}(\hat{\theta})} \right] \quad (2.9)$$

El intervalo de confianza dado en (2.9) contendrá al verdadero parámetro  $\theta$  para una proporción de muestras, extraídas bajo el mismo diseño, de aproximadamente  $1 - \alpha$ , si se verifican las siguientes condiciones:

1. La distribución del estimador  $\hat{\theta}$  en el muestreo es aproximadamente normal con esperanza  $\theta$  y varianza  $\mathbb{V}(\hat{\theta})$ . Esta condición es equivalente a afirmar que el Teorema Central del Límite se puede aplicar a  $\hat{\theta}$ .
2. Existe un estimador consistente,  $\hat{\mathbb{V}}(\hat{\theta})$ , para  $\mathbb{V}(\hat{\theta})$ .

Bajo estas condiciones, si consideramos la variable aleatoria

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\mathbb{V}}(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{\mathbb{V}(\hat{\theta})}} \cdot \left( \frac{\mathbb{V}(\hat{\theta})}{\hat{\mathbb{V}}(\hat{\theta})} \right)^{1/2},$$

es fácil ver que es una variable aleatoria con distribución aproximadamente  $N(0, 1)$ , por lo que puede justificarse el uso del intervalo confianza dado en (2.9). Nótese que si  $\mathbb{V}(\hat{\theta})$  es conocida se usará el intervalo de confianza

$$IC_{\theta}(s) = \left[ \hat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(\hat{\theta})} \right]. \quad (2.10)$$

## Bibliografía

- Groves, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Hansen, M.H. y W.N. Hurwitz (1943). "On the theory of sampling from finite populations". En: *Ann. Math. Statist.* 14, págs. 333-362.
- Horvitz, D.G. y D.J. Thompson (1952). "A generalization of sampling without replacement from a finite universe". En: *Journal of the American Statistical Association* 47, págs. 663-685.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
- Sen, A.R. (1953). "On the estimate of the variance in sampling with varying probabilities". En: *Journal of the Indian Society of Agricultural Statistics* 5, págs. 119-127.
- Tillé, Y. (2006). *Sampling algorithms*. Springer.
- Yates, F. y P.M. Grundy (1953). "Selection without replacement from within strata with probability proportional to size". En: *Journal of the Royal Statistical Society B* 15, págs. 253-261.

## Tema 3

### **Muestreo de Bernoulli: definición, estimadores, varianza y estimador de la varianza. Muestreo aleatorio simple: sin y con reemplazamiento: definición, estimadores, varianza y estimador de la varianza.**

Este tema está elaborado como una adaptación casi literal en español del capítulo 3 de la siguiente bibliografía:

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

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.

### **3.1 Introducción**

El muestreo de Bernoulli y el muestreo aleatorio simple son dos tipos de diseños de muestreo directo de elementos. La selección de la muestra bajo este tipo de diseños requiere la existencia de un marco que identifique las unidades de muestreo, que son los elementos de la población.

Así pues, los diseños de muestreo directo de elementos presentan dos características importantes:

1. Existe un marco de muestreo que identifica a cada elemento de la población.
2. Las unidades de muestreo son precisamente los elementos de la población.

Existen otros diseños de muestreo directo de elementos, como el muestreo sistemático, el muestreo de Poisson, el muestreo con probabilidades proporcionales al tamaño o el muestreo estratificado. Estos diseños serán estudiados en los temas sucesivos.

Considérese una población constituida por  $N$  elementos  $\{u_1, \dots, u_N\}$ , que se denota por  $U = \{1, \dots, N\}$  y supóngase que el parámetro de interés es el total poblacional de la *variable de estudio*  $y$ .

Tanto el muestreo de Bernoulli como el muestreo aleatorio simple pertenecen a la categoría de diseños denominados *diseños de muestreo con probabilidades iguales*. La característica común de este tipo de diseños es que las probabilidades de inclusión de primer orden  $\pi_k$  son iguales, esto es  $\pi_k = \text{constante}$ , para todo  $k = 1, \dots, N$ .  $\pi_k$  representa la probabilidad de que el elemento  $u_k$  pertenezca a la muestra. Se denotará  $u_k \in s$  o bien  $k \in s$  para indicar que el elemento  $k$  pertenece a la muestra aleatoria.

Supongamos que desea obtenerse información sobre la variable ‘ingreso del hogar’. A esta variable de interés se la denomina *variable de estudio*. Podríamos estar interesados por ejemplo en obtener información sobre el total de ingresos de los hogares de la población, esto es,

$$Y_U = \sum_{k \in U} y_k$$

En este tema se obtendrán los estimadores convenientes de  $Y_U$  para los diseños de muestreo de Bernoulli y muestreo aleatorio simple, así como las expresiones de la varianza de estos estimadores y sus estimadores.

## 3.2 Muestreo de Bernoulli

### 3.2.1 Definición

El muestreo de Bernoulli es un diseño de muestreo con probabilidades iguales que produce muestras de tamaño variable. Denotamos el muestreo de Bernoulli como diseño *Brs*.

Las probabilidades de inclusión de primer y segundo orden de todos los elementos de la población son iguales, ya que los indicadores de pertenencia a la muestra  $I_k(s) = I_k$  son variables aleatorias independientes e idénticamente distribuidas con distribución Bernoulli de parámetro  $\pi$ :

$$\mathbb{P}(I_k = 1) = \pi, \quad \mathbb{P}(I_k = 0) = 1 - \pi$$

por lo que se tiene

$$\mathbb{E}(I_k) = \pi_k = \pi, \quad k = 1, \dots, N.$$

Las probabilidades de inclusión de segundo orden son:

$$\pi_{kl} = \mathbb{E}(I_k \cdot I_l) = \mathbb{E}(I_k) \cdot \mathbb{E}(I_l) = \pi^2, \quad \text{para todo } k \neq l,$$

$$\text{donde } I_k \cdot I_l = \begin{cases} 1, & \text{si } u_k, u_l \in s, \\ 0, & \text{si } u_k, u_l \notin s. \end{cases}$$

Cualquier posible muestra es una secuencia de  $N$  experimentos independientes. Si denotamos por  $n_s$  la variable aleatoria que representa el número de elementos seleccionados en la muestra aleatoria  $s$ , entonces la probabilidad de seleccionar dicha muestra es  $\pi^{n_s} \cdot (1 - \pi)^{N - n_s}$ . Por tanto, el diseño muestral de Bernoulli es:

$$p(s) = \pi^{n_s} \cdot (1 - \pi)^{N - n_s}, \quad \text{para todo } s \in \Omega,$$

donde  $\Omega$  es el espacio muestral, el conjunto de posibles muestras.

La probabilidad de que el tamaño de la muestra aleatoria sea  $n_s$  es:

$$\mathbb{P}(n_s) = \binom{N}{n_s} \cdot \pi^{n_s} \cdot (1 - \pi)^{N-n_s}, \quad n_s = 0, 1, \dots, N.$$

$n_s$  tiene distribución binomial de parámetros  $N$  y  $\pi$ , por lo que la esperanza y la varianza son, respectivamente:

$$\begin{aligned} \mathbb{E}(n_s) &= N \cdot \pi, \\ \mathbb{V}(n_s) &= N \cdot \pi \cdot (1 - \pi). \end{aligned}$$

Usando la distribución normal como aproximación de la binomial, podemos evaluar el rango de valores probables de  $n_s$ , de forma que  $n_s$  está contenido dentro del intervalo

$$\left[ N \cdot \pi \pm z_{\alpha/2} \cdot [N \cdot \pi \cdot (1 - \pi)]^{1/2} \right]$$

con una probabilidad de aproximadamente  $1 - \alpha$ , siendo  $z_{\alpha/2}$  el valor que verifica  $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$  donde  $Z$  es una variable aleatoria con distribución  $N(0, 1)$ .

A continuación se describe un posible procedimiento para seleccionar una muestra de Bernoulli. Un procedimiento de selección de muestras probabilísticas es un algoritmo que consiste en la realización secuencial de experimentos aleatorizados que produce como resultado un elemento seleccionado en la muestra tras cada experimento o bien la inclusión o exclusión en la muestra de cada elemento del marco.

*Ejemplo 18.* Un ejemplo de implementación del diseño *Brs* es el siguiente:

- Se consideran  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ .
- Sea  $\pi$  una constante fijada tal que  $0 < \pi < 1$ , la selección o no selección de la unidad  $u_k$  en la muestra se determina de la siguiente forma: si  $\varepsilon_k < \pi$ , el elemento  $u_k$  es seleccionado en la muestra, en caso contrario no.

■

Del algoritmo de selección anterior se puede deducir que la probabilidad de selección de cualquier elemento de la población es  $\pi$ :

$$\mathbb{P}(u_k \in s) = \mathbb{P}(\varepsilon_k < \pi) = \pi, \quad k = 1 \dots, N.$$

Dado que las realizaciones son independientes, la elección del elemento  $k$  y el elemento  $l$  en la muestra para cualesquiera  $k \neq l$  son dos eventos independientes. Por tanto:

$$\mathbb{P}(u_k, u_l \in s) = \mathbb{P}(u_k \in s) \cdot \mathbb{P}(u_l \in s) = \pi^2.$$

### 3.2.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador de Horvitz-Thompson (HT) como estimador del total poblacional  $Y_U$ . Veamos a continuación cuál es el estimador y la varianza bajo el diseño  $Brs$ . Para demostrar los resultados se hará uso de las expresiones obtenidas en el Tema 2.

#### Proposición 7

Bajo el diseño muestral de Bernoulli, el estimador HT de  $Y_U$  viene dado por

$$\hat{Y}_U^{\text{HT}} = \frac{1}{\pi} \sum_{k \in s} y_k. \quad (3.1)$$

La varianza del estimador de Horvitz-Thompon,  $\hat{Y}_U^{\text{HT}}$  es

$$\mathbb{V}_{Brs}(\hat{Y}_U^{\text{HT}}) = \left(\frac{1}{\pi} - 1\right) \sum_{k \in U} y_k^2. \quad (3.2)$$

Un estimador insesgado de la varianza del estimador es

$$\hat{\mathbb{V}}_{Brs}(\hat{Y}_U^{\text{HT}}) = \frac{1}{\pi} \cdot \left(\frac{1}{\pi} - 1\right) \sum_{k \in s} y_k^2. \quad (3.3)$$

#### Demostración 7

$$\begin{aligned} \hat{Y}_U^{\text{HT}} &= \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{\pi} \sum_{k \in s} y_k \\ \mathbb{V}_{Brs}(\hat{Y}_U^{\text{HT}}) &= \sum_{k \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l) \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \cdot (1 - \pi_k) + \underbrace{\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l)}_0 \\ &= \sum_{k \in U} \frac{y_k^2}{\pi} \cdot (1 - \pi) = \left(\frac{1}{\pi} - 1\right) \sum_{k \in U} y_k^2. \end{aligned}$$

donde se ha usado que  $\pi_{kl} = \pi^2$  y  $\pi_k = \pi$ ,  $\forall k \in U$ .

Por último, se demuestra la insesgadez del estimador:

$$\mathbb{E}[\hat{\mathbb{V}}_{Brs}(\hat{Y}_U^{\text{HT}})] = \mathbb{E}\left[\frac{1}{\pi} \cdot \left(\frac{1}{\pi} - 1\right) \sum_{k \in s} y_k^2\right] =$$

$$\begin{aligned}
&= \frac{1}{\pi} \cdot \left( \frac{1}{\pi} - 1 \right) \sum_{k \in U} y_k^2 \cdot \mathbb{E}(I_k) = \\
&= \left( \frac{1}{\pi} - 1 \right) \sum_{k \in U} y_k^2
\end{aligned}$$

*Ejemplo 19.* Un profesor de universidad que debe corregir 600 exámenes escritos desea tener una idea preliminar de la tasa de aprobados. Para tal propósito, decide realizar una primera corrección a mano de un número más reducido de exámenes. La selección de la muestra sigue el siguiente algoritmo: para cada examen, el profesor lanza un dado de seis caras, si sale un 6, corrige el examen, mientras que si obtiene cualquier otro valor, no lo corrige. Es fácil observar que este algoritmo muestral implementa el diseño de Bernoulli con  $\pi = \frac{1}{6}$ . Calculemos ahora una estimación del número de estudiantes que han aprobado el examen basada en el estimador HT, suponiendo que se han seleccionado 90 exámenes y 60 de ellos han aprobado el examen. La variable de estudio  $y$  se define como una variable dicotómica,  $y_k = 1$  si el estudiante  $u_k$  aprueba el examen e  $y_k = 0$  si no lo aprueba. Entonces:

$$\hat{Y}_U^{\text{HT}} = \frac{1}{\pi} \sum_{k \in s} y_k = 6 \cdot 60 = 360.$$

Usando (3.3), la estimación de la varianza es:

$$\hat{V}_{Brs}(\hat{Y}_U^{\text{HT}}) = \frac{1}{\pi} \cdot \left( \frac{1}{\pi} - 1 \right) \cdot \sum_{k \in s} y_k^2 = 6 \cdot 5 \cdot 60 = 1800.$$

■

Los diseños que producen muestras de tamaño variable, como el diseño muestral de Bernoulli, presentan algunas desventajas.

En primer lugar, el hecho de no conocer al principio el tamaño de muestra supone cierta pérdida de control sobre el trabajo de campo. Por ejemplo, el presupuesto disponible podría sobrepasarse considerablemente si el tamaño de la muestra obtenido supera notablemente el tamaño esperado.

Por otra parte, es importante seleccionar un estimador cuya varianza no esté penalizada por el hecho de que el tamaño de muestra es variable. El estimador HT es a menudo ineficiente bajo el diseño muestral de Bernoulli, ya que la variabilidad del tamaño de muestra penaliza la varianza del estimador  $\hat{Y}_U^{\text{HT}}$ . Este inconveniente puede ser subsanado si se elige un estimador adecuado.

Es por ello que se propone a continuación el siguiente estimador de  $Y_U$  mejorado,  $\hat{Y}_U^{\text{Rat}}$ ,

$$\hat{Y}_U^{\text{Rat}} = \frac{N}{n_s} \cdot \sum_{k \in s} y_k$$

La aparición del tamaño de muestra  $n_s$  del denominador en la definición del estimador tiene el efecto de reducir la parte de la variabilidad que es debida a la variación del tamaño muestral.

El estimador  $\hat{Y}_U^{Rat}$  puede escribirse en función del estimador HT como

$$\hat{Y}_U^{Rat} = \frac{N}{n_s} \cdot \sum_{k \in s} y_k = \frac{n}{n_s} \cdot \frac{N}{n} \sum_{k \in s} y_k = \frac{n}{n_s} \cdot \hat{Y}_U^{HT}$$

donde  $n = N \cdot \pi = \mathbb{E}(n_s)$  es el tamaño de muestra esperado.

**Comentario 2.** La varianza del estimador  $\hat{Y}_U^{HT}$  bajo el diseño *Brs*, dada en (3.2), puede escribirse también como

$$\mathbb{V}_{Brs}(\hat{Y}_U^{HT}) = N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2 \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right] \quad (3.4)$$

donde  $S_{yU}^2$  es la cuasivarianza poblacional de la variable  $y$  y  $CV(Y_U)$  es el coeficiente de variación poblacional.

$$S_{yU}^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y}_U)^2 = \frac{1}{N-1} \cdot \left( \sum_{k \in U} y_k^2 - N \cdot \bar{y}_U^2 \right)$$

Se puede llegar a:

$$\begin{aligned} \sum_{k \in U} y_k^2 &= (N-1) \cdot S_{yU}^2 + N \bar{y}_U^2 \\ &= (N-1) \cdot S_{yU}^2 + N \cdot S_{yU}^2 \cdot \frac{1}{\left( \frac{\bar{y}_U}{S_{yU}} \right)^2} \\ &= N \cdot S_{yU}^2 \cdot \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right]. \end{aligned}$$

Por tanto:

$$\begin{aligned} \mathbb{V}_{Brs}(\hat{Y}_U^{HT}) &= \left( \frac{1}{\pi} - 1 \right) \sum_{k \in U} y_k^2 = \left( \frac{N}{n} - 1 \right) \cdot N \cdot S_{yU}^2 \cdot \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right] \\ &= N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2 \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right]. \end{aligned}$$

■

**Comentario 3.** Se puede demostrar que la varianza aproximada *AV* del estimador  $\hat{Y}_U^{Rat}$  viene dada por

$$AV(\hat{Y}_U^{Rat}) = N \cdot \left( \frac{1}{\pi} - 1 \right) \cdot S_{yU}^2 = N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2$$



Véase el Tema 3<sup>1</sup> del bloque “Producción Estadística Oficial: Métodos Avanzados” del grupo de materias específicas, sobre la estimación de razón.



A continuación se realizará la comparación del estimador HT y el estimador mejorado bajo el diseño *Brs*. Consideremos, por un lado, la estrategia  $e_1$  compuesta por el diseño muestral de Bernoulli  $p(s)$  y el estimador HT  $\hat{Y}_U^{HT}$ , y, por otro lado, la estrategia  $e_2$  formada por el mismo diseño muestral  $p(s)$  y el estimador mejorado  $\hat{Y}_U^{Rat}$ . El efecto de diseño de la estrategia  $e_1$  comparado con la estrategia  $e_2$  es:

$$\frac{V_{Brs}(\hat{Y}_U^{HT})}{V_{Brs}(\hat{Y}_U^{Rat})} = \frac{N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot S_{yU}^2 \left[1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2}\right]}{N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot S_{yU}^2} = 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2}$$

Para un tamaño de población considerable, el efecto de diseño es aproximadamente  $1 + \frac{1}{CV(Y_U)^2}$ , por lo que se puede observar que la ganancia de eficiencia obtenida usando el estimador  $\hat{Y}_U^{Rat}$  es mayor cuanto menor es el valor del coeficiente de variación poblacional de  $y$ .

*Ejemplo 20.* Continuando con el ejemplo anterior, calculemos ahora una estimación del número de alumnos aprobados basada en el estimador mejorado  $\hat{Y}_U^{Rat}$ :

$$\hat{Y}_U^{Rat} = \frac{N}{n_s} \sum_{k \in s} y_k = \frac{600}{90} \cdot 60 = 400.$$

Se obtiene una estimación de 400 alumnos aprobados, respecto a los 360 estimados usando el estimador de Horvitz-Thompson. No podemos decir qué estimación es mejor, dado que no se tiene aún el número real de aprobados hasta que los 600 exámenes sean corregidos. Únicamente podría decirse que la estimación puntual de 400 es mejor en el sentido de que proviene de un estimador con menor variabilidad.

Suponiendo que el número de estudiantes aprobados es de 390, podemos obtener el coeficiente de variación poblacional:

$$CV(Y_U) = \frac{S_{yU}}{\bar{y}_U} = \left(\frac{N}{N-1} \cdot \frac{Q}{P}\right)^{1/2} = \left(\frac{600}{599} \cdot \frac{210/600}{390/600}\right)^{1/2} = \left(\frac{600}{599} \cdot \frac{7}{13}\right)^{1/2}$$

donde  $P$  representa la proporción poblacional de estudiantes aprobados y  $Q = 1 - P$ .

El efecto de diseño es

$$\frac{V_{Brs}(\hat{Y}_U^{HT})}{V_{Brs}(\hat{Y}_U^{Rat})} = 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \approx 2,85.$$

<sup>1</sup>Tema 3. Introducción a problemas de estimación complejos. El efecto del sesgo en intervalos de confianza de las estimaciones. Consistencia e insesgadez asintótica. La técnica de linealización de Taylor para la estimación de la varianza. Estimador de una razón: varianza y sesgo.

Se puede observar que el estimador mejorado es más preciso. ■

### 3.3 Muestreo aleatorio simple sin reemplazamiento

#### 3.3.1 Definición

El muestreo aleatorio simple sin reemplazamiento o sin reposición es un diseño de muestreo con probabilidades iguales que produce muestras de tamaño fijo. Denotamos el muestreo aleatorio simple sin reemplazamiento como diseño *srswor*.

El diseño *srswor* consiste en seleccionar un subconjunto de  $n$  elementos de la población, no repetidos, donde cada elemento tiene la misma probabilidad de pertenecer a la muestra. Téngase en cuenta que en este diseño no hay reemplazamiento de las unidades seleccionadas previamente. Por otra parte, muestras con los mismos elementos pero con un orden distinto se consideran iguales. Así, la cardinalidad del espacio muestral es  $\binom{N}{n}$  y todas las muestras tienen probabilidad igual a  $\frac{1}{\binom{N}{n}}$ .

El diseño muestral es:

$$p(s) = \frac{1}{\binom{N}{n}}, \text{ para todo } s \in \Omega.$$

Bajo el diseño *srswor*, las probabilidades de inclusión de primer y segundo orden de todos los elementos de la población son iguales. En efecto:

$$\begin{aligned} \pi_k &= \mathbb{P}(u_k \text{ aparezca una vez en la muestra y las } n-1 \text{ unidades} \\ &\quad \text{restantes que forman parte de la muestra no sean } u_k) = \\ &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}; \quad k = 1, \dots, N. \\ \pi_{kl} &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n \cdot (n-1)}{N \cdot (N-1)}; \quad k \neq l = 1, \dots, N. \end{aligned}$$

A continuación se presentan varios procedimientos para implementar el diseño *srswor* y obtener como resultado una muestra aleatoria simple sin reemplazamiento.

*Ejemplo 21.* Consideremos el siguiente algoritmo secuencial de selección de la muestra, que consiste en realizar  $n$  experimentos aleatorizados (en este caso extracciones) donde el resultado es la selección de un elemento en la muestra.

1. Se selecciona un elemento de los  $N$  posibles con igual probabilidad:  $\frac{1}{N}$ .
2. Se selecciona un segundo elemento de entre los  $N-1$  restantes con igual probabilidad:  $\frac{1}{N-1}$ .

3. ∴

- n. Se selecciona un elemento de entre los  $N - n + 1$  restantes con igual probabilidad:  $\frac{1}{N-n+1}$ .

■

Sin embargo, cuando el tamaño de la población es grande suele ser más conveniente usar mecanismos donde el resultado de cada experimento sea la inclusión o exclusión en la muestra del elemento. Veamos algunos ejemplos de algoritmos de este tipo.

*Ejemplo 22.* Este mecanismo de selección de la muestra es debido a [Fan, Muller y Rezucha 1962](#).

- Se consideran  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ .
- Sea  $\frac{n}{N}$  una constante fijada, la selección o no selección de la primera unidad  $u_1$  en la muestra se determina de la siguiente forma: si  $\varepsilon_1 < \frac{n}{N}$ , el elemento  $u_1$  es seleccionado en la muestra, en caso contrario no.
- Para los siguientes elementos de la población,  $k = 2, \dots, N$ , el elemento es seleccionado si

$$\varepsilon_k < \frac{n - n_k}{N - k + 1}$$

donde  $n_k$  es el tamaño de la muestra hasta ese momento, es decir, el número de elementos que han resultado seleccionados en la muestra de entre los primeros  $k - 1$  elementos para los que ya se ha realizado el experimento.

- El procedimiento termina cuando  $n_k = n$ .

■

Nótese que para implementar el esquema descrito en el Ejemplo 22 el tamaño de la población  $N$  debe ser conocido. En caso de no ser conocido, es necesario realizar una pasada preliminar a lo largo del marco muestral para determinar  $N$ . En la práctica puede ocurrir que el tamaño de la población de nuestro estudio sea desconocido, por ejemplo si la población de interés está referida a un subconjunto de los elementos del marco.

*Ejemplo 23.* [McLeod y Bellhouse 1983](#) proponen un método sencillo de selección de una muestra aleatoria simple de tamaño  $n$ , que no requiere del conocimiento previo del tamaño  $N$ .

- Se seleccionan los primeros  $n$  elementos en la muestra inicial:  $k_1, k_2, \dots, k_n$ .
- Para cualquier elemento  $k > n$ , se considera una realización de una variable aleatoria con distribución uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon_k$ . Se calcula  $j = 1 + [\varepsilon_k \cdot k]$ , donde  $[\cdot]$  denota la parte entera. Si  $j \leq n$  entonces el elemento  $j$  de la muestra actual es reemplazado por el elemento  $k$  de la población.

- El algoritmo termina cuando se ha realizado el experimento para todos los elementos.

Por tanto, al finalizar el algoritmo  $k$  contiene el valor del tamaño poblacional  $N$ .

■

*Ejemplo 24.* Otro esquema para seleccionar una muestra aleatoria simple es el siguiente:

- Se consideran  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo  $(0, 1) : \varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ .
- Los valores obtenidos se ordenan de menor a mayor

$$\varepsilon_{(k_1)} < \varepsilon_{(k_2)} < \dots < \varepsilon_{(k_N)}$$

donde  $\varepsilon_{(k_j)}$  es el  $j$ -ésimo valor más pequeño de los  $N$  valores obtenidos.

- La muestra de tamaño  $n$  estará compuesta por los  $n$  primeros elementos que se corresponden precisamente con aquellos para los que se obtuvieron los valores  $\varepsilon_k$  más pequeños  $\varepsilon_{(k_1)}, \varepsilon_{(k_2)}, \dots, \varepsilon_{(k_n)}$ .

■

El esquema de selección de muestras descrito en el Ejemplo 24 presenta la ventaja de que permite obtener de forma simultánea varias muestras aleatorias simples no solapadas, es decir, donde cualquier elemento de la población solo puede haber sido seleccionado en una muestra. La segunda muestra estaría formada por los siguientes  $n$  elementos que se corresponden con los que siguientes valores  $\varepsilon_k$  más bajos. Esta muestra no está solapada con la primera. Y así sucesivamente.

Las muestras sin solapamientos son deseables, por ejemplo, cuando es necesario realizar varias encuestas diferentes sobre la misma población en un corto espacio de tiempo. Esto es un beneficio, ya que así se reduce la carga del informante.

### 3.3.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador HT como estimador del total poblacional  $Y_U$ . Veamos a continuación cuál es el estimador y la varianza bajo el diseño *srswor*. Para demostrar los resultados se hará uso de las expresiones obtenidas en el Tema 2.

**Proposición 8**

El estimador HT es

$$\hat{Y}_U^{\text{HT}} = \frac{N}{n} \cdot \sum_{k \in s} y_k.$$

La varianza del estimador viene dada por

$$\mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}}) = N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2 = N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2, \quad (3.5)$$

donde  $f = \frac{n}{N}$  se denomina fracción de muestreo.

**Demostración 8**

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{\frac{n}{N}} = \frac{N}{n} \cdot \sum_{k \in s} y_k$$

La expresión de la varianza de  $\hat{Y}_U^{\text{HT}}$  es:

$$\begin{aligned} \mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}}) &= \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \cdot (1 - \pi_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \pi_l) \\ &= \sum_{k \in U} \frac{y_k^2}{\frac{n}{N}} \cdot \left( 1 - \frac{n}{N} \right) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\frac{n}{N}} \cdot \frac{y_l}{\frac{n}{N}} \cdot \left[ \frac{n \cdot (n-1)}{N \cdot (N-1)} - \frac{n}{N} \cdot \frac{n}{N} \right] \\ &= \frac{N-n}{n} \cdot \left[ \sum_{k \in U} y_k^2 - \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k \cdot y_l}{N-1} \right]. \end{aligned}$$

Teniendo en cuenta que la varianza es invariante ante un cambio de origen, se tiene:

$$\mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}}) = \frac{N-n}{n} \cdot \left[ \sum_{k \in U} (y_k - \bar{y}_U)^2 - \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{(y_k - \bar{y}_U) \cdot (y_l - \bar{y}_U)}{N-1} \right].$$

Recordando que  $\sum_{k \in U} (y_k - \bar{y}_U) = 0$ , se tiene

$$\left( \sum_{k \in U} (y_k - \bar{y}_U) \right)^2 = \sum_{k \in U} (y_k - \bar{y}_U)^2 + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} (y_k - \bar{y}_U) \cdot (y_l - \bar{y}_U) = 0.$$

De esta forma, la expresión de  $\mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}})$  queda:

$$\begin{aligned} \mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}}) &= \frac{N-n}{n} \cdot \left[ \sum_{k \in U} (y_k - \bar{y}_U)^2 + \frac{1}{N-1} \cdot \sum_{k \in U} (y_k - \bar{y}_U)^2 \right] = \\ &= \frac{N-n}{n} \cdot \left( 1 + \frac{1}{N-1} \right) \cdot \sum_{k \in U} (y_k - \bar{y}_U)^2 = \\ &= N^2 \cdot (1-f) \cdot \frac{S_{yU}^2}{n}. \end{aligned}$$

### Proposición 9

Bajo el diseño del muestreo aleatorio simple sin reemplazamiento, un estimador insesgado de la cuasivarianza poblacional es la cuasivarianza muestral:

$$S_{ys}^2 = \frac{1}{n-1} \cdot \sum_{k \in s} (y_k - \bar{y}_s)^2.$$

### Demostración 9

$$\begin{aligned} \mathbb{E}[S_{ys}^2] &= \frac{1}{n-1} \cdot \mathbb{E} \left[ \sum_{k \in s} (y_k^2 - 2 \cdot y_k \cdot \bar{y}_s + \bar{y}_s^2) \right] = \frac{1}{n-1} \cdot \mathbb{E} \left[ \sum_{k \in s} y_k^2 - n \cdot \bar{y}_s^2 \right] \\ &= \frac{n}{n-1} \cdot \mathbb{E} \left[ \sum_{k \in U} \frac{y_k^2}{n} \cdot I_k - \bar{y}_s^2 \right] = \frac{n}{n-1} \cdot \left[ \sum_{k \in U} \frac{y_k^2}{n} \cdot \underbrace{\mathbb{E}[I_k]}_{\frac{n}{N}} - \mathbb{E}[\bar{y}_s^2] \right] \\ &= \frac{n}{n-1} \cdot \left[ \frac{1}{N} \cdot \sum_{k \in U} y_k^2 - \bar{y}_U^2 - \mathbb{V}_{srswor}(\bar{y}_s) \right] = \frac{n}{n-1} \cdot [\sigma_{yU}^2 - \mathbb{V}(\bar{y}_s)]. \end{aligned}$$

Usando (3.5) se tiene:

$$\begin{aligned} \mathbb{E}[S_{ys}^2] &= \frac{n}{n-1} \cdot \left[ \sigma_{yU}^2 - \frac{1-f}{n} \cdot S_{yU}^2 \right] \\ &= \frac{n}{n-1} \cdot \left[ \sigma_{yU}^2 - \frac{1-f}{n} \cdot S_{yU}^2 \right] \end{aligned}$$

$$\underbrace{S_{yU}^2 = \frac{N}{N-1} \cdot \sigma_{yU}^2}_{= \frac{N}{N-1} \cdot \sigma_{yU}^2} = \frac{N}{N-1} \cdot \sigma_{yU}^2 = S_{yU}^2.$$

**Corolario 10**

Un estimador insesgado de la varianza del estimador de Horvitz-Thompson es

$$\hat{V}_{srswor}(\hat{Y}_U^{HT}) = N^2 \cdot \frac{1-f}{n} \cdot S_{ys}^2 \quad (3.6)$$

**Demostración 10**

Al resultado se llega de directamente utilizando que la cuasivarianza muestral es un estimador insesgado de la cuasivarianza poblacional.

**Comentario 4.** La expresión debida a [Yates y Grundy 1953](#) y [Sen 1953](#) para el estimador de la varianza del estimador HT dada en (2.6), aplicada al diseño *srswor* coincide con (3.6).

$$\begin{aligned} \hat{V}_{srswor}(\hat{Y}_U^{HT}) &= -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \\ &= -\frac{1}{2} \cdot \left[ -\frac{1-f}{n-1} \right] \cdot \frac{1}{f^2} \sum_{k \in s} \sum_{l \in s} (y_k - y_l)^2. \end{aligned}$$

Teniendo en cuenta que la varianza es invariante ante un cambio de origen, se tiene:

$$\begin{aligned} \hat{V}_{srswor}(\hat{Y}_U^{HT}) &= -\frac{1}{2} \cdot \left[ -\frac{1-f}{n-1} \right] \cdot \frac{1}{f^2} \sum_{k \in s} \sum_{l \in s} [(y_k - \bar{y}_s) - (y_l - \bar{y}_s)]^2 \\ &= \frac{1}{2} \cdot \frac{1-f}{n-1} \cdot \frac{1}{f^2} \left[ \sum_{k \in s} \sum_{l \in s} 2(y_k - \bar{y}_s)^2 - 2 \sum_{k \in s} \sum_{l \in s} (y_k - \bar{y}_s)(y_l - \bar{y}_s) \right] \\ &= \frac{1-f}{n-1} \cdot \frac{1}{f^2} \sum_{k \in s} n \cdot (y_k - \bar{y}_s)^2 = \frac{1-f}{n-1} \cdot \frac{1}{f^2} \cdot n \cdot (n-1) \cdot S_{ys}^2 \\ &= N^2 \cdot \frac{1-f}{n} \cdot S_{ys}^2. \end{aligned}$$

■

**Ejemplo 25.** Se desea estimar el consumo mensual de folios de los trabajadores de una empresa. Para ello se decide realizar un muestreo aleatorio simple sin reemplazamiento. Se seleccionan 200 de los 1000 trabajadores que hay en la empresa, obteniéndose un consumo medio de 60 folios por trabajador, y una cuasivarianza muestral de 25. La estimación del consumo mensual de folios a partir del uso del estimador de Horvitz-

Thompson es

$$\hat{Y}_U^{\text{HT}} = \frac{N}{n} \sum_{k \in s} y_k = N \cdot \bar{y}_s = 1000 \cdot 60 = 60000.$$

La estimación de la varianza del estimador es

$$\hat{\mathbb{V}}_{\text{srswor}} \left( \hat{Y}_U^{\text{HT}} \right) = N^2 \cdot (1 - f) \cdot \frac{S_{ys}^2}{n} = 1000^2 \cdot \left( 1 - \frac{200}{1000} \right) \cdot \frac{25}{200} = 100000.$$

Luego, la estimación del error de muestreo vendrá dado, por definición, por la raíz cuadrada de 100000.

Además de obtener una estimación puntual del consumo mensual de folios, podríamos dar un intervalo de confianza al 95 %. Suponiendo la normalidad de la distribución del estimador, el intervalo de confianza quedaría en este caso:

$$IC_{Y_U}(s) = \left[ \hat{Y}_U^{\text{HT}} \pm t_{n-1; \alpha/2} \cdot \hat{\sigma}_{\text{srswor}} \left( \hat{Y}_U^{\text{HT}} \right) \right] = \left[ 60000 \pm 1,96 \cdot \sqrt{100000} \right],$$

donde  $t_{n-1; \alpha/2}$  es el valor que verifica  $\mathbb{P}(T > t_{n-1; \alpha/2}) = \alpha/2$  considerando  $T$  una variable aleatoria con distribución  $t$  de Student con  $n - 1$  grados de libertad.

■

### Estimación de la media poblacional

Un estimador insesgado de la media poblacional  $\bar{y}_U$  se obtiene dividiendo el estimador de Horvitz-Thompson para el caso del total por el tamaño poblacional  $N$ :

$$\hat{\bar{y}}_U^{\text{HT}} = \frac{1}{N} \cdot \hat{Y}_U^{\text{HT}} = \frac{1}{n} \cdot \sum_{k \in s} y_k = \bar{y}_s$$

En efecto:

$$\mathbb{E} \left[ \hat{\bar{y}}_U^{\text{HT}} \right] = \mathbb{E} \left[ \frac{1}{N} \cdot \hat{Y}_U^{\text{HT}} \right] = \frac{1}{N} \cdot Y_U = \bar{y}_U.$$

De esta forma, la varianza y el estimador de la varianza pueden obtenerse de forma sencilla:

$$\begin{aligned} \mathbb{V}_{\text{srswor}} \left( \hat{\bar{y}}_U^{\text{HT}} \right) &= \frac{1}{N^2} \cdot \mathbb{V} \left[ \hat{Y}_U^{\text{HT}} \right] = \frac{1-f}{n} \cdot S_{yU}^2, \\ \hat{\mathbb{V}}_{\text{srswor}} \left( \hat{\bar{y}}_U^{\text{HT}} \right) &= \frac{1-f}{n} \cdot S_{ys}^2. \end{aligned}$$



### 3.3.3 Estimación en dominios

En la mayoría de encuestas se suele estar interesado en realizar estimaciones no solo para la población completa  $U$ , sino también para determinadas subpoblaciones, denominadas *dominios de estudio* o simplemente *dominios*. Por ejemplo, podríamos tener interés en estimar el número de personas desempleadas por grupos de edad, sexo, ocupación o bien por regiones del país. La estimación realizada en determinadas subpoblaciones se conoce como *estimación en dominios*.

Sea  $U_d$  el dominio o subpoblación de interés, se define  $N_d$  como el tamaño de  $U_d$ , esto es, el número de elementos de la población que pertenecen al dominio  $U_d$  y, por otra parte,  $P_d$ , la proporción de elementos de la población  $U$  que pertenecen a  $U_d$ .

- **Estimación del tamaño absoluto y relativo de un dominio**

Suponiendo conocido  $N$  y desconocido  $N_d$ , se desea estimar el valor de  $N_d$  y  $P_d$ . Para ello, se define la variable  $z_d$ , que toma los valores

$$z_{dk} = \begin{cases} 1, & \text{si } k \in U_d, \\ 0, & \text{si } k \notin U_d, \end{cases}$$

para todo  $k \in 1, \dots, N$ .

El total de la variable  $z_d$  es el número de elementos que pertenecen a  $U_d$  y la media de la variable  $z_d$  es la proporción de elementos que pertenecen a  $U_d$ :

$$\begin{aligned} \sum_{k \in U} z_{dk} &= N_d \\ \bar{z}_{dU} = \sum_{k \in U} \frac{z_{dk}}{N} &= \frac{N_d}{N} = P_d. \end{aligned}$$

Sea  $s$  la muestra aleatoria seleccionada de tamaño  $n$  con  $n_d = \sum_{k \in s} z_{dk}$  el número de elementos de la muestra que pertenecen al dominio y  $p_d = \frac{n_d}{n}$  la proporción de elementos de la muestra que pertenecen a  $U_d$ .

A continuación se aplicarán los resultados de la Proposición 8 para obtener el estimador HT de  $N_d$  y la varianza del estimador

$$\hat{N}_d = N \cdot p_d.$$

Tomando  $Q_d = 1 - P_d$ , la varianza del estimador viene dada por:

$$\mathbb{V}_{srswor}(\hat{N}_d) = N^2 \cdot (1 - f) \cdot \frac{S_{z_dU}^2}{n} = N^2 \cdot \frac{N - n}{N - 1} \cdot \frac{\sigma_{z_dU}^2}{n} = N^2 \cdot \frac{N - n}{N - 1} \cdot \frac{P_d \cdot Q_d}{n}$$

ya que

$$\sigma_{z_dU}^2 = \frac{1}{N} \cdot \sum_{k \in U} (z_{dk} - P_d)^2 = \frac{1}{N} \cdot \left[ \sum_{k \in U} z_{dk}^2 - 2 \cdot P_d \sum_{k \in U} z_{dk} + N \cdot P_d^2 \right]$$

$$\begin{aligned}
&= \frac{1}{N} \cdot \left[ \sum_{k \in U} z_{dk} - 2 \cdot N \cdot P_d^2 + N \cdot P_d^2 \right] = \frac{1}{N} \cdot [N \cdot P_d - N \cdot P_d^2] \\
&= P_d \cdot (1 - P_d) = P_d \cdot Q_d.
\end{aligned}$$

Usando (3.6), un estimador insesgado de la varianza es:

$$\widehat{V}_{srswor}(\widehat{N}_d) = N^2 \cdot (1 - f) \cdot \frac{S_{ys}^2}{n} = N^2 \cdot (1 - f) \cdot \frac{p_d \cdot q_d}{n - 1},$$

ya que  $S_{ys}^2 = \frac{n}{n-1} \cdot p_d \cdot q_d$ .

De forma análoga, para el caso de la proporción de individuos que pertenecen al dominio, un estimador insesgado para  $P_d$  es

$$\widehat{P}_d = p_d.$$

La varianza y el estimador de la varianza se pueden calcular a partir de las expresiones vistas para el estimador de  $N_d$ :

$$\begin{aligned}
V_{srswor}(\widehat{P}_d) &= \frac{1}{N^2} \cdot V_{srswor}(\widehat{N}_d) = \frac{N - n}{N - 1} \cdot \frac{P_d \cdot Q_d}{n}, \\
\widehat{V}_{srswor}(\widehat{P}_d) &= \frac{1}{N^2} \cdot \widehat{V}_{srswor}(\widehat{N}_d) = (1 - f) \cdot \frac{p_d \cdot q_d}{n - 1}.
\end{aligned}$$

**Comentario 5.** Estos resultados son importantes ya que permiten la estimación de la proporción poblacional y el total poblacional de los elementos de la población que pertenecen a un determinado dominio, como se muestra en el ejemplo siguiente. ■

*Ejemplo 26.* En un territorio donde existen 1500 colegios, se desea conocer la opinión de estos acerca de un nuevo proyecto educativo que se pretende implantar. Para ello, se selecciona una muestra aleatoria simple sin reemplazamiento de 300 colegios, obteniéndose la siguiente información muestral:

A favor	En contra	En blanco
225	50	25

Sea  $z_{dk} = \begin{cases} 1, & \text{si el colegio está a favor del nuevo proyecto} \\ 0, & \text{en otro caso} \end{cases}$

Una estimación de la proporción de colegios a favor del proyecto educativo, usando el estimador HT, es

$$\widehat{P}_d = p_d = \frac{225}{300} = 0,75.$$

Mientras que la estimación del número de colegios a favor del proyecto educativo es  $\hat{N} = 1500 \cdot 0,75 = 1125$ .

Un intervalo de confianza al 95 % aproximadamente para la proporción viene dado por

$$\left[ \hat{P}_d \pm z_{\alpha/2} \cdot \hat{\sigma}_{srswor}(\hat{P}_d) \right] = [0,75 \pm 1,96 \cdot 0,0224],$$

$$\text{ya que } \hat{\sigma}_{srswor}(\hat{P}_d) = \sqrt{\left(1 - \frac{300}{1500}\right) \cdot \frac{0,75 \cdot 0,25}{299}} \approx 0,0224.$$

Por otra parte,  $z_{\alpha/2}$  es el valor que verifica  $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$  donde  $Z$  es una variable aleatoria con distribución  $N(0, 1)$ .

■

• **Estimación del total y la media de un dominio cuando el tamaño del dominio es desconocido.**

Supongamos que estamos interesados en estimar el total de ingresos de los hogares con familia numerosa, esto es,

$$Y_{U_d} = \sum_{k \in U_d} y_k.$$

Se define una nueva variable,  $y_d$ , tal que

$$y_{dk} = \begin{cases} y_k, & \text{si } k \in U_d, \\ 0, & \text{otro caso.} \end{cases}$$

El total de ingresos de los hogares con familia numerosa puede expresarse en función de esta nueva variable teniendo en cuenta todos los hogares de la población y, de esta manera, poder hacer uso de los resultados vistos en la Proposición 8 y el Corolario 10:

$$Y_{U_d} = \sum_{k \in U} y_{dk}$$

El estimador HT es

$$\hat{Y}_{U_d}^{\text{HT}} = \sum_{k \in s} \frac{y_{dk}}{\pi_k} = \frac{N}{n} \sum_{k \in s} y_{dk} = \frac{N}{n} \sum_{k \in s_d} y_k,$$

donde  $s_d$  es el subconjunto de la muestra aleatoria  $s$  que pertenece al dominio  $U_d$ .

La varianza de  $\hat{Y}_{U_d}^{\text{HT}}$  viene dada por

$$\begin{aligned} \mathbb{V}_{srswor}(\hat{Y}_{U_d}^{\text{HT}}) &= \\ &= N^2 \cdot \frac{1-f}{n} \cdot S_{y_d U}^2 = N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \cdot \left[ \sum_{k \in U} y_{dk}^2 - N \cdot \left( \frac{1}{N} \sum_{k \in U} y_{dk} \right)^2 \right] = \end{aligned}$$

$$= N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \cdot \left[ \sum_{k \in U_d} y_k^2 - \frac{1}{N} \cdot \left( \sum_{k \in U_d} y_k \right)^2 \right]$$

Un estimador insesgado de la varianza es

$$\begin{aligned} \widehat{\mathbb{V}}_{srswor} \left( \widehat{Y}_{U_d}^{\text{HT}} \right) &= \\ &= N^2 \cdot \frac{1-f}{n} \cdot S_{y_d S}^2 = N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \cdot \left[ \sum_{k \in s} y_{dk}^2 - n \cdot \left( \frac{1}{n} \sum_{k \in s} y_{dk} \right)^2 \right] = \\ &= N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \cdot \left[ \sum_{k \in s_d} y_k^2 - \frac{1}{n} \cdot \left( \sum_{k \in s_d} y_k \right)^2 \right] \end{aligned}$$

**Comentario 6.** En numerosas ocasiones el tamaño del dominio  $N_d$  es desconocido. Sin embargo, si  $N_d$  fuese conocido, uno normalmente preferiría usar el estimador alternativo

$$\widehat{Y}_U^{\text{Rat}} = \frac{N_d}{n_d} \sum_{k \in s_d} y_k = N_d \cdot \bar{y}_{s_d},$$

Nótese que  $n_d$  es una variable aleatoria. La varianza y estimador de la varianza de este tipo de estimadores se estudiarán en el tema 3<sup>2</sup> del bloque “Producción Estadística Oficial: Métodos Avanzados” del grupo de materias específicas. ■

### 3.3.4 Comparación del muestreo aleatorio simple sin reemplazamiento y el muestreo de Bernoulli

Consideremos, por un lado, la estrategia  $e_1$  compuesta por el diseño muestral de Bernoulli y el estimador de Horvitz-Thompson y, por otra parte, la estrategia  $e_2$  formada por el diseño del muestreo aleatorio simple sin reemplazamiento y el estimador de Horvitz-Thompson también.

Usando (3.4), la varianza del estimador HT bajo el diseño  $Brs$  es

$$\begin{aligned} \mathbb{V}_{Brs} \left( \widehat{Y}_U^{\text{HT}} \right) &= N \cdot \left( \frac{1-\pi}{\pi} \right) \cdot S_{yU}^2 \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right] \\ &= N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2 \left[ 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2} \right] \end{aligned}$$

donde  $n = N \cdot \pi = \mathbb{E}(n_s)$  es el tamaño de muestra esperado.

Por otro lado, tenemos que la varianza del estimador HT bajo el diseño  $srswor$  es

$$\mathbb{V}_{srswor} \left( \widehat{Y}_U^{\text{HT}} \right) = N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2.$$

<sup>2</sup>Tema 3. Introducción a problemas de estimación complejos. El efecto del sesgo en intervalos de confianza de las estimaciones. Consistencia e insesgadez asintótica. La técnica de linealización de Taylor para la estimación de la varianza. Estimador de una razón: varianza y sesgo.

Para obtener una comparación justa del muestreo de Bernoulli con respecto al muestreo aleatorio simple sin reemplazamiento que produce una muestra de tamaño  $n$  fijo, se fija el tamaño de muestra esperado en el diseño de Bernoulli, de tal forma que  $n = N \cdot \pi$ .

El efecto de diseño de la estrategia  $e_1$  comparado con la estrategia  $e_2$  es:

$$\begin{aligned} \frac{V_{Brs}(\hat{Y}_U^{HT})}{V_{srswr}(\hat{Y}_U^{HT})} &= \frac{N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot S_{yU}^2 \left[1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2}\right]}{N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2} = \\ &= \frac{\frac{N-n}{n \cdot N}}{\frac{N-n}{N} \cdot \frac{1}{n}} \cdot \left[1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2}\right] = \\ &= 1 - \frac{1}{N} + \frac{1}{CV(Y_U)^2}. \end{aligned}$$

Para un tamaño de población considerable, el efecto de diseño es aproximadamente  $1 + \frac{1}{CV(Y_U)^2}$ . Se puede observar que el muestreo de Bernoulli es a menudo considerablemente menos preciso que el muestreo aleatorio simple cuando se usa el estimador de Horvitz-Thompson. Esta pérdida de precisión es debida a la variabilidad del tamaño de la muestra en el diseño de Bernoulli.

### 3.4 Muestreo aleatorio simple con reemplazamiento

#### 3.4.1 Definición

Un diseño muestral con reemplazamiento permite la selección de muestras con unidades repetidas, a diferencia de los esquemas sin reemplazamiento, que producen muestras con todos sus elementos distintos.

El muestreo aleatorio simple con reemplazamiento, también denominado muestreo aleatorio simple con reposición, es un diseño de muestreo con probabilidades iguales que produce muestras de tamaño fijo. Denotamos el muestreo aleatorio simple con reemplazamiento como diseño *srswr*.

El diseño *srswr* consiste en seleccionar de forma independiente  $n$  elementos de la población de tamaño  $N$  con una probabilidad igual de extracción de  $p_k = \frac{1}{N}$ ,  $\forall k = 1, \dots, N$ . Téngase en cuenta que en este diseño hay reemplazamiento de cada elemento extraído, es decir, elementos seleccionados en la muestra pueden ser elegidos de nuevo en la siguiente extracción.

*Ejemplo 27.* El siguiente procedimiento es una posible implementación del diseño *srswr*. Consiste en seleccionar secuencialmente y de forma independiente elementos de la población hasta obtener un subconjunto de  $n$  unidades:  $k_1, k_2, \dots, k_n$ . En cada extracción, todos los elementos de la población tienen la misma probabilidad de extracción:  $p_k = \frac{1}{N}$ .

Cada vez que un elemento es elegido, se vuelve a reponer, de forma que en cada extracción los  $N$  elementos de la población son susceptibles de ser seleccionados.

■

Sea  $os = (k_1, \dots, k_n)$  la muestra ordenada resultante del algoritmo muestral, donde se tiene información sobre el orden de extracción y el número de veces que cada elemento aparece en la muestra. Para nuestro propósito, la información sobre el orden de extracción no es de interés, pues consideraremos como la misma muestra aquellas que contengan los mismos elementos. Por tanto, la cardinalidad del espacio muestral es una combinación con repetición de  $N$  elementos tomados de  $n$  en  $n$ .

Se define la variable aleatoria  $m_k$  como el número de veces que aparece el elemento  $k$  en la muestra, cuya distribución es binomial de parámetros  $n$  y  $\frac{1}{N}$ , por lo que la esperanza y la varianza son, respectivamente:

$$\begin{aligned}\mathbb{E}(m_k) &= n \cdot \frac{1}{N}, \\ \mathbb{V}(m_k) &= n \cdot \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)\end{aligned}$$

La probabilidad de seleccionar una muestra viene dada por el modelo multinomial:

$$\mathbb{P}(e_1 = m_1, \dots, e_N = m_N) = \frac{n!}{m_1! \cdot \dots \cdot m_N!} \cdot \left(\frac{1}{N}\right)^n.$$

En el diseño *srswr* las probabilidades de inclusión de primer y segundo orden de todos los elementos de la población son iguales.

$$\begin{aligned}\pi_k &= 1 - \left(1 - \frac{1}{N}\right)^n, \quad k = 1, \dots, N, \\ \pi_{kl} &= 1 - 2 \cdot \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n, \quad k \neq l = 1, \dots, N.\end{aligned}$$

### 3.4.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador de Hansen y Hurwitz (HH) como estimador del total poblacional  $Y_U$ . Veamos a continuación cuál es el estimador y la varianza bajo el diseño *srswr*. Para demostrar los resultados se hará uso de expresiones obtenidas en el Tema 2.

**Proposición 11**

El estimador de Hansen-Hurwitz es

$$\hat{Y}_U^{HH} = \frac{N}{n} \cdot \sum_{i=1}^n y_{k_i}.$$

La varianza del estimador viene dada por

$$\mathbb{V}_{srswr}(\hat{Y}_U^{HH}) = N \cdot (N - 1) \cdot \frac{S_{yU}^2}{n} = N^2 \cdot \frac{\sigma_{yU}^2}{n}. \quad (3.7)$$

**Demostración 11**

$$\hat{Y}_U^{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{\frac{1}{N}} = \frac{N}{n} \cdot \sum_{i=1}^n y_{k_i}$$

En cuanto a la varianza del estimador, usando la expresión (2.7), se llega a

$$\begin{aligned} \mathbb{V}_{srswr}(\hat{Y}_U^{HH}) &= \frac{1}{n} \sum_{k \in U} \left( \frac{y_k}{p_k} - Y_U \right)^2 \cdot p_k = \frac{1}{n} \sum_{k \in U} \left( \frac{y_k}{\frac{1}{N}} - Y_U \right)^2 \cdot \frac{1}{N} \\ &= \frac{1}{n} \sum_{k \in U} \frac{(N \cdot y_k - N \cdot \bar{y}_U)^2}{N} = \frac{N^2}{n} \sum_{k \in U} \frac{(y_k - \bar{y}_U)^2}{N} \\ &= \frac{N^2}{n} \cdot \sigma_{yU}^2. \end{aligned}$$

**Comentario 7.** En el diseño de muestreo aleatorio simple los estimadores de Horvitz-Thompson y Hansen y Hurwitz coinciden. ■

**Proposición 12**

Bajo el diseño del muestreo aleatorio simple con reemplazamiento, un estimador insesgado de la varianza poblacional es la cuasivarianza muestral:

$$S_{yos}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (y_{k_i} - \bar{y}_{os})^2$$

donde  $\bar{y}_{os} = \frac{1}{n} \sum_{i=1}^n y_{k_i}$ .

**Demostración 12**

En la Demostración 9 se llegó al siguiente resultado:

$$\mathbb{E} [S_{yos}^2] = \frac{n}{n-1} \cdot [\sigma_{yU}^2 - \mathbb{V}(\bar{y}_{os})].$$

Usando (3.7) se tiene lo que se pretendía demostrar

$$\mathbb{E} [S_{yos}^2] = \frac{n}{n-1} \cdot \left[ \sigma_{yU}^2 - \frac{\sigma_{yU}^2}{n} \right] = \sigma_{yU}^2.$$

**Corolario 13**

Un estimador insesgado de la varianza es

$$\hat{\mathbb{V}}(\hat{Y}_U^{HH}) = N^2 \cdot \frac{S_{yos}^2}{n}.$$

**Demostración 13**

Al resultado se llega de forma obvia utilizando que la cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

Dada una *muestra ordenada*  $os = (k_1, \dots, k_n)$  podemos formar el conjunto de elementos distintos que existen, que denotamos por  $s^*$ :

$$s^* = \{k : k = k_i \text{ para algún } i, i = 1, \dots, n\}$$

A partir de este conjunto podemos formar el siguiente estimador insesgado alternativo al estimador de Hansen y Hurwitz:

$$\hat{Y}_U^{Rat} = \frac{N}{n_{s^*}} \cdot \sum_{k \in s^*} y_k,$$

donde  $n_{s^*} \leq n$  es el número de elementos distintos que hay en  $os$ . Se puede demostrar que el estimador  $\hat{Y}_U^{Rat}$ , que tiene únicamente en cuenta las unidades distintas de la muestra, presenta una varianza que no excede la varianza del estimador de Hansen y Hurwitz, el cual considera todas las unidades de la muestra. Este resultado es debido a [Basu 1958](#) y [Raj y Khamis 1958](#). Véase ([Cassel, Särndal y Wretman 1977](#)) para una demostración.

### 3.4.3 Comparación del muestreo aleatorio simple sin y con reemplazamiento

Consideremos por un lado la estrategia  $e_1$  compuesta por el diseño *srswr* y el estimador HH y, por otra parte, la estrategia  $e_2$  formada por el diseño *srswor* y el estimador HT. El



efecto de diseño es:

$$\frac{V_{srswr}(\hat{Y}_U^{HH})}{V_{srswor}(\hat{Y}_U^{HT})} = \frac{N \cdot (N-1) \cdot \frac{S_{yU}^2}{n}}{N^2 \cdot (1 - \frac{n}{N}) \cdot \frac{S_{yU}^2}{n}} = \frac{N-1}{N \cdot (1 - \frac{n}{N})} = \frac{N-1}{N-n} \geq 1, \text{ para todo } n \geq 1.$$

Se obtiene que el muestreo sin reemplazamiento es más eficiente (salvo si el tamaño de la muestra es  $n = 1$ ).

*Ejemplo 28.* Volvamos al contexto del Ejemplo 25, donde se estaba interesado en estimar el consumo mensual de folios de los trabajadores de una empresa. Nos preguntamos ahora cuál sería el tamaño de muestra necesario para garantizar la misma precisión (error de muestreo) usando el diseño *srswr* que la obtenida con el diseño *srswor*, suponiendo que la cuasivarianza poblacional es 30.

Por un lado, la varianza del estimador HT es

$$\mathbb{V}_{srswor}(\hat{Y}_U^{HT}) = N^2 \cdot (1-f) \cdot \frac{S_{yU}^2}{n} = 1000^2 \cdot \left(1 - \frac{200}{1000}\right) \cdot \frac{30}{200} = 120000.$$

La varianza del estimador HH es

$$\mathbb{V}_{srswr}(\hat{Y}_U^{HH}) = N^2 \cdot \frac{\sigma_{yU}^2}{n'}.$$

Dado que  $\sigma_{yU}^2 = \frac{N-1}{N} \cdot S_{yU}^2$ , el tamaño de muestra necesario puede calcularse como sigue:

$$n' = 1000^2 \cdot \frac{\frac{999}{1000} \cdot 30}{120000} = 249,75.$$

Como cabía esperar, se necesitará un tamaño de muestra mayor, 250. ■

## Bibliografía

- Basu, D. (1958). "On sampling with and without replacement". En: *Sankhya* 20, págs. 287-294.
- Cassel, C.-M., C.-E. Särndal y J.H. Wretman (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.
- Fan, C.T., M.E. Muller e I. Rezucha (1962). "Development of sampling plans by using sequential (item by item) techniques and digital computers". En: *Journal of the American Statistical Association* 57, págs. 387-402.
- McLeod, A.I. y D.R. Bellhouse (1983). "A convenient algorithm for drawing a simple random sample". En: *Applied Statistics* 32, págs. 182-184.
- Raj, D. y S. H. Khamis (1958). "Some remarks on sampling with replacement". En: *Annals of Mathematical Statistics* 29, págs. 550-557.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

- Sen, A.R. (1953). "On the estimate of the variance in sampling with varying probabilities". En: *Journal of the Indian Society of Agricultural Statistics* 5, págs. 119-127.
- Yates, F. y P.M. Grundy (1953). "Selection without replacement from within strata with probability proportional to size". En: *Journal of the Royal Statistical Society B* 15, págs. 253-261.

## Tema 4

### Estimación insesgada en diseños muestrales sobre unidades elementales II. Muestreo sistemático: definición, estimadores, varianza del estimador. Problemática de la estimación de la varianza. La eficiencia del muestreo sistemático.

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

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

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.

#### 4.1 Introducción

En los Temas 3 a 6 de 'Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas' se discuten estimaciones insesgadas para el muestreo directo de elementos. Hay dos características del muestreo directo de elementos:

- (i) existe un marco de muestreo que permite la identificación de cada uno de los elementos de la población;
- (ii) en la selección de la muestra, los elementos poblacionales son unidades muestrales.

Se pueden considerar los siguiente diseños de muestreo directo de elementos:

- i. Muestreo de Bernoulli ( $Brs$ ), que se ve en el Tema 3.
- ii. Muestreo aleatorio simple, sin reemplazamiento ( $srs_{wor}$ ) y con reemplazamiento ( $srs_{wr}$ ), que también se ven en el Tema 3.
- iii. Muestreo sistemático ( $sys$ , del inglés *systematic*), que se ve en este tema.

- iv. Muestreo de Poisson ( $Prs$ ), que se ve en el Tema 5.
- v. Muestreo con probabilidades proporcionales al tamaño, sin reemplazamiento ( $ppswor$ ) y con reemplazamiento ( $ppswr$ , del inglés *probability proportional to size without/with replacement*), que se ven en el Tema 5.
- vi. Muestreo estratificado ( $strs$ , del inglés *stratified random sampling*), que se ve en el Tema 6.

La mayoría de estos diseños son muy utilizados en la práctica. Los diseños  $Brs$  y  $Prs$  dan lugar a tamaños muestrales aleatorios. Sin embargo, estos diseños son muy adecuados para ilustrar algunas ideas básicas en el muestreo de encuestas y son útiles también como modelos para mecanismos de respuestas para la estimación en presencia de falta de respuesta.

El parámetro de principal interés en este tema es el total poblacional de la variable de estudio  $y$ , es decir,  $Y_U = \sum_{k \in U} y_k$ . En este tema estudiaremos una estimación *insesgada* de  $Y_U$ , centrándonos en el estimador de Horvitz-Thompson,

$$\hat{Y}_U^{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}. \quad (4.1)$$

## 4.2 Muestreo sistemático: definición, estimadores, varianza del estimador

El muestreo sistemático se refiere a un conjunto de procedimientos que ofrecen varias ventajas prácticas, en particular su simplicidad de ejecución. Nos centraremos en el muestreo sistemático en su forma básica. Un primer elemento se selecciona aleatoriamente y con igual probabilidad entre los primeros  $a$  elementos del marco muestral. El entero positivo  $a$  se fija previamente y se llama *intervalo de muestreo*<sup>1</sup>. No hace falta ninguna otra extracción aleatoria. El resto de la muestra se selecciona sistemáticamente tomando cada uno de los siguientes  $a$  elementos a continuación, hasta el final del marco. Por tanto, hay solo  $a$  muestras posibles, cada una con una probabilidad  $\frac{1}{a}$  de ser seleccionada. La simplicidad de una única selección aleatoria es una gran ventaja. Es fácil, por ejemplo, para un entrevistador seleccionar una muestra sistemática mientras se encuentra en tareas de campo.

Para una definición formal de este tipo de muestreo sistemático, sea  $a$  el intervalo de muestreo fijo y sea  $n$  la parte entera de  $\frac{N}{a}$ , donde  $N$  es el tamaño poblacional. Entonces

$$N = na + c,$$

donde el entero  $c$  verifica  $0 \leq c < a$ . Si  $c = 0$ , el tamaño muestral  $n$  será seleccionado mediante el procedimiento que presentamos a continuación. Si  $c > 0$ , el tamaño muestral será  $n$  o  $n + 1$ . La selección, que se puede ver como una lista secuencial, es de la siguiente forma:

---

<sup>1</sup>*Sampling interval.*

- i. Se selecciona con igual probabilidad  $\frac{1}{a}$  un entero aleatoriamente, digamos  $r$ , entre 1 y  $a$  (incluido).
- ii. La muestra seleccionada está compuesta por

$$s = \{k : k = r + (j - 1)a \leq N; j = 1, 2, \dots, n_s\} = s_r, \quad (4.2)$$

donde el tamaño muestral  $n_s$  es  $n + 1$  (cuando  $r \leq c$ ) o  $n$  (cuando  $c < r \leq a$ ).

El entero  $r$  se llama *arranque aleatorio*.

*Ejemplo 29.* Supongamos un profesor universitario que está corrigiendo 600 exámenes escritos y decide hacer una evaluación preliminar de la nota de corte para aprobar. Tiene una variedad de diseños muestrales para seleccionar una muestra. Puede usar el muestreo sistemático de la siguiente forma: seleccionar el primer alumno, tirar el dado una vez, supongamos que sale '2'. La muestra sistemática está entonces totalmente determinada; está constituida por los estudiantes 2, 8, 14, ..., y 596. El tamaño muestral es 100 y  $a = 6$ ,  $c = 0$ . ■

**Comentario 8.** A veces, el muestreo sistemático se utiliza como sustituto del muestreo aleatorio simple cuando no se dispone de un marco poblacional o cuando éste último tiene un orden más o menos aleatorio. Por ejemplo, siguiendo con el caso del profesor universitario, supongamos que quiere elegir una muestra de 45 estudiantes de una lista de 45000 que estudian en su universidad. El intervalo de muestreo  $k$  es 1000. Supongamos que el entero aleatorio elegido es 597. Entonces los estudiantes numerados como 597, 1597, 2597, ..., 44597 formarían la muestra.

Si los nombres de los estudiantes están en orden alfabético, es probable que obtengamos una muestra con un comportamiento similar al de una muestra aleatoria simple; es poco probable que la posición alfabética de una persona quede asociada con la característica de interés. Sin embargo, el muestreo sistemático no es igual al muestreo aleatorio simple ya que no se verifica la propiedad de que cada grupo posible de  $n$  elementos tenga la misma probabilidad de ser la muestra elegida. En el ejemplo del párrafo anterior, es imposible que los estudiantes 345 y 346 aparezcan en la muestra. Desde el punto de vista técnico, el muestreo sistemático es una forma de muestreo por conglomerados, que se verá en el Tema 7.

La mayor parte del tiempo, una muestra sistemática proporciona resultados comparables con los de una muestra aleatoria simple, y los métodos de muestreo aleatorio simple se pueden usar en el análisis. Si la población tiene un orden aleatorio, la muestra sistemática será muy similar a una muestra aleatoria simple. Se puede pensar que la población está mezclada. Kennedy leía una de cada 50 cartas de las treinta mil que llegaban semanalmente a la Casa Blanca. Esta muestra sistemática se comportaba como una muestra aleatoria. Cabe señalar que Kennedy era consciente de que las cartas que leía, aunque representativas de las cartas que recibía en la Casa Blanca, no eran

representativas de la opinión pública.

El muestreo sistemático no proporciona, necesariamente, una muestra representativa <sup>2</sup> si la lista de unidades de la población tiene algún orden periódico o cíclico. Por ejemplo, si los hombres y las mujeres se alternan en la lista y  $k$  es par, la muestra sistemática sólo tendrá hombres o mujeres, lo que no puede considerarse una muestra representativa. En los estudios ecológicos realizados en terrenos agrícolas puede haber una topografía accidentada que produzca un patrón periódico de vegetación. Si un esquema de muestreo sistemático sigue el mismo ciclo, la muestra no se comportará como una muestra aleatoria simple.

Por otro lado, algunas poblaciones tienen un orden creciente o decreciente. Una lista de cantidades a cobrar puede estar ordenada de mayor a menor cantidad. En este caso, las estimaciones de una muestra sistemática podrían tener una varianza menor (pero inestable) en relación con las estimaciones respectivas de la muestra aleatoria simple. Una muestra sistemática de una lista ordenada de cantidades por cobrar debe tener algunas cantidades grandes y otras pequeñas. Es posible que una muestra aleatoria simple sólo contenga cantidades pequeñas o sólo cantidades grandes, de modo que haya más variabilidad entre las medias muestrales de todas las muestras aleatorias simples posibles que entre las medias muestrales de todas las muestras sistemáticas posibles.

En el muestreo sistemático debemos de seguir contando con un marco muestral y tener cuidado al definir la población objetivo. El muestreo de uno de cada 20 estudiantes que entran en la biblioteca no nos dará una muestra representativa del conjunto de estudiantes. Sin embargo, el muestreo de una de cada 10 personas que salen de un avión probablemente dará una muestra representativa de las personas de ese vuelo. El marco de muestreo para los pasajeros del avión no aparece de forma explícita, pero existe de cualquier modo. ■

El conjunto de posibles muestras bajo diseño muestral sistemático, denotado por  $\mathcal{S}_{sys}$  consiste en  $a$  conjuntos diferentes (que no se solapan) que se puede obtener de la forma de la ecuación (4.2), concretamente,

$$\mathcal{S}_{sys} = \{s_1, \dots, s_r, \dots, s_a\}.$$

Habitualmente, esto representa un número extremadamente bajo de muestras posibles comparado con, por ejemplo, el muestreo aleatorio simple sin reemplazamiento.

El diseño muestral sistemático, que denotaremos por  $sys$ , viene dado por

$$p(s) = \begin{cases} \frac{1}{a} & \text{si } s \in \mathcal{S}_{sys}, \\ 0 & \text{para cualquier otra muestra } s. \end{cases}$$

<sup>2</sup>se entiende representativo como la propiedad deseable de una estrategia muestral (diseño + estimador) que conduce a estimaciones acuradas

Para construir el estimador de Horvitz-Thompson y su varianza necesitamos las probabilidades de inclusión de primer y segundo orden. Como cada elemento  $k$  pertenece a uno y solo una de las  $a$  muestras sistemáticas igualmente probables, tenemos, para cada  $k \in U$ ,

$$\pi_k = \frac{1}{a},$$

mientras que, para cada  $k \neq l \in U$ ,

$$\pi_{kl} = \begin{cases} \frac{1}{a} & \text{si } s \text{ y } l \text{ pertenecen a la misma muestra } s, \\ 0 & \text{en cualquier otro caso.} \end{cases} \quad (4.3)$$

La propiedad deseable de que todos los  $\pi_{kl} > 0$  (véase el Tema 2 de este mismo bloque) no se verifica, por tanto.

**Comentario 9.** No hay solapamiento entre cualesquiera dos muestras y las  $a$  muestras juntas forman la población total  $U$ . Es decir,

$$s_1, \dots, s_r, \dots, s_a$$

representa una partición disjunta de  $U$  en subpoblaciones. Es decir,

$$U = \bigcup_{r=1}^a s_r$$

y podemos escribir  $Y_U = \sum_{k \in U} y_k$  como

$$Y_U = \sum_{r=1}^a Y_{s_r} \quad \text{con} \quad Y_{s_r} = \sum_{k \in s_r} y_k.$$

El diseño muestral sistemático puede, por tanto, describirse como una selección aleatoria con igual probabilidad de una de las  $a$  subpoblaciones. En la subpoblación seleccionada se encuesta a todos sus miembros. La Tabla 4.1 ilustra la situación para el caso en el  $c = 0$ .

	Muestra, $s$				
	$s_1$	$\dots$	$s_r$	$\dots$	$s_a$
Valores de $y$	$y_1$	$\dots$	$y_r$	$\dots$	$y_a$
	$y_{a+1}$	$\dots$	$y_{a+r}$	$\dots$	$y_{2a}$
	$\vdots$		$\vdots$		$\vdots$
Total muestral	$\frac{y_{(n-1)a+1}}{Y_{s_1}}$	$\dots$	$\frac{y_{(n-1)a+r}}{Y_{s_r}}$	$\dots$	$\frac{y_N}{Y_{s_a}}$

Tabla 4.1: Muestreo sistemático en el caso  $c = 0$ .



Por tanto, como consecuencia del resultado general sobre el estimador HT y su varianza (véase el Tema 2 de este bloque), tenemos el siguiente resultado:

#### Teorema 14

Bajo el diseño muestral sistemático, con intervalo de muestreo  $a$ , el estimador HT del total poblacional  $Y_U = \sum_{k \in U} y_k$  toma la forma

$$\hat{Y}_U^{\text{HT}} = aY_s, \quad (4.4)$$

donde  $Y_s = \sum_{k \in s} y_k$  es el total muestral de  $y$  y  $s$  es uno de los elementos del conjunto de posibles muestras  $\{s_1, \dots, s_r, \dots, s_a\}$ , con  $s_r$  definido por (4.2):

$$s = \{k : k = r + (j - 1)a \leq N; j = 1, 2, \dots, n_s\} = s_r.$$

La varianza viene dada por

$$\mathbb{V}_{sys}(\hat{Y}_U^{\text{HT}}) = a \sum_{r=1}^a (Y_{s_r} - \bar{Y}_U)^2, \quad (4.5)$$

donde  $\bar{Y}_U = \sum_{r=1}^a \frac{Y_{s_r}}{a} = \frac{Y_U}{a}$  es una media de los totales muestrales  $Y_{s_r} = \sum_{k \in s_r} y_k$ .

#### Demostración 14

Puesto que  $\pi_k = \frac{1}{a}$ ,  $k \in U$ , tenemos

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = aY_s.$$

Para determinar la varianza, usamos el resultado general sobre el estimador HT su varianza, junto con (4.3), para obtener

$$\begin{aligned} \mathbb{V}_{sys}(\hat{Y}_U^{\text{HT}}) &= \sum_{k \in U} \sum_{l \in U} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - Y_U^2 \\ &= a \sum_{r=1}^a \left\{ \sum_{k \in s_r} \sum_{l \in s_r} y_k y_l \right\} - Y_U^2 \\ &= a \sum_{r=1}^a Y_{s_r}^2 - Y_U^2 = a \sum_{r=1}^a (Y_{s_r} - \bar{Y}_U)^2. \end{aligned}$$



La varianza también se puede escribir como

$$\mathbb{V}_{sys}(\hat{Y}_U^{HT}) = a(a-1)S_Y^2,$$

donde

$$S_Y^2 = \frac{1}{a-1} \sum_{r=1}^a (Y_{s_r} - \bar{Y})^2$$

es la varianza de los totales muestrales. La varianza será pequeña si los totales muestrales son casi iguales.

**Comentario 10.** Como la condición  $\pi_{kl} > 0$  para todos los  $k \neq l$  no se verifica, no debería usarse el estimador HT de la varianza de  $\hat{Y}_U^{HT}$ , esto es,

$$\hat{\mathbb{V}}^{HT}[\hat{Y}_U^{HT}] = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

La fórmula da, en este caso, un resultado absurdo. La estimación de la varianza en el muestreo sistemático se discute en la Sección 4.4. ■

### 4.2.1 El control del tamaño muestral

A partir de la definición del diseño sistemático se sigue que si  $c = 0$ , entonces  $N = na$  y todas las  $a$  posibles muestras tienen el mismo tamaño  $n$ . Si, por otro lado,  $c > 0$ , el tamaño muestral será  $n+1$  (si  $r \leq c$ ) o  $n$  (si  $r > c$ ). También está claro que imponer a  $a$  el requisito de que sea un entero positivo llevará, en casos extremos, a tamaños muestrales que pueden diferir considerablemente de lo que es deseable.

Por ejemplo, supongamos que  $N = 149$  y que el tamaño muestral deseado es 60. Entonces la elección de  $a = 2$  da un tamaño muestral de 74 o 75, mientras que  $a = 3$  da un tamaño muestral de 49 o 50. Los tamaños entre 50 y 74 son imposibles. Existen distintos métodos para gestionar este problema. A continuación se presentan dos. Obviamente, si  $N$  es muy grande comparado con  $n$  el problema es leve.

#### *i. Método de intervalo fraccionario*

Este método permite un valor fraccionario de  $a$ . Sea  $a = \frac{N}{n}$ , donde  $n$  es el tamaño muestral deseado. Se selecciona un número aleatorio  $\xi$  con la distribución uniforme en el intervalo  $(0, a)$ . La muestra seleccionada estará constituida entonces por aquellos elementos  $k$  para los que

$$k-1 < \xi + (j-1)a \leq k; \quad j = 1, \dots, n.$$

El método es equivalente a seleccionar, con igual probabilidad  $\frac{1}{N}$ , un entero aleatorio  $r$  entre 1 y  $N$  (incluido) y seleccionar el elemento  $k$  para el que

$$(k-1)n < r + (j-1)N \leq kn; \quad j = 1, \dots, n.$$

Por ejemplo, el elemento  $k = 1$  se selecciona si  $r$  verifica  $0 < r \leq n$ , lo que ocurre con probabilidad  $\frac{n}{N}$ .

De forma similar, cualquier elemento  $k$  tendrá probabilidad  $\pi_k = \frac{1}{a} = \frac{n}{N}$  de ser elegido y cada muestra posible tendrá exactamente tamaño  $n$ .

#### ii. Método de muestreo sistemático circular

En este método, el marco se diseña circularmente, es decir, el último elemento ( $k = N$ ) va seguido del primero ( $k = 1$ ), y así sucesivamente. Se selecciona un número aleatorio  $r$  entre 1 y  $N$  (incluido) con igual probabilidad. Sea  $a$  el entero más cercano a  $\frac{N}{n}$ . Entonces la muestra consiste en  $k$  elementos tales que, para  $j = 1, \dots, n$ ,

$$k = r + (j - 1)a \quad \text{si} \quad r + (j - 1)a \leq N$$

o

$$k = r + (j - 1)a - N \quad \text{si} \quad r + (j - 1)a > N$$

Como en el método del intervalo fraccionario, cada muestra tendrá tamaño  $n$  y  $\pi_k = \frac{n}{N}$  para cada  $k$ .

**Comentario 11.** Para calcular la varianza del estimador HT para los métodos del intervalo fraccionario y circular, debemos primero calcular las probabilidades  $\pi_{kl}$ . Esto exige atención especial a que las muestras posibles no son necesariamente disjuntas por parejas en estos dos métodos. En el caso  $c = 0$ , los tres métodos de muestra sistemático presentados en esta sección son equivalentes. Cuando  $N$  es grande comparado con  $n$  las diferencias entre los tres métodos es pequeña. ■

### 4.3 La eficiencia del muestreo sistemático

La varianza (4.5) está cerca de 0 si los totales muestrales  $Y_{s_r}$  son aproximadamente iguales. ¿Cuándo ocurre esto? Recordamos que las muestras  $s_r$  se formaron tomando sistemáticamente cada  $a$ -ésimo elemento. Si la ordenación de los  $N$  elementos de la población es tal que las muestras sistemáticas resultantes tienen aproximadamente el mismo total de  $y$ , la varianza será pequeña. En otras palabras, la eficiencia del muestreo sistemático depende en gran medida de la ordenación de los  $N$  elementos sobre los cuales se realiza el muestreo sistemático.

Veamos ahora la eficiencia del muestreo sistemático en función de la ordenación de la población. Por simplicidad, consideremos el caso en que  $N = an$ , donde  $a$  es un entero. Por el resultado dado en (4.2), el estimador HT construido con la muestra  $s = s_r$  viene dado por

$$\hat{Y}_U^{\text{HT}} = N \sum_{k \in s_r} \frac{y_k}{n} = N \bar{y}_{s_r}, \quad (4.6)$$

cuya varianza es

$$\mathbb{V}_{sys}(\hat{Y}_U^{HT}) = \frac{N^2}{a} \sum_{r=1}^a (\bar{y}_{s_r} - \bar{y}_U)^2, \quad (4.7)$$

con  $\bar{y}_U = \sum_{k \in U} \frac{y_k}{N}$ , que es la media poblacional.

La varianza mostrada en la ecuación (4.7) es cercana a 0 si todas las medias muestrales  $\bar{y}_{s_r}$  son aproximadamente iguales. Para analizar en profundidad la varianza en (4.7), hay que tener en cuenta que

$$\sum_{k \in U} (\bar{y}_{s_r} - \bar{y}_U)^2 = \sum_{r=1}^a \sum_{k \in s_r} (\bar{y}_k - \bar{y}_{s_r})^2 + \sum_{r=1}^a n(\bar{y}_{s_r} - \bar{y}_U)^2, \quad (4.8)$$

es decir, la variación total en la población se puede descomponer en la variación dentro de las muestras sistemáticas<sup>3</sup> y la variación entre las muestras sistemáticas<sup>4</sup>, como en el análisis de la varianza (ANOVA). Esto se puede escribir como

$$SST = SSW + SSB, \quad (4.9)$$

donde  $SS$  denota la suma de cuadrados:  $T$  total;  $W$  dentro (*within*) y  $B$  entre (*between*). Para una población dada,  $SST = (N - 1)S_{yU}^2$  es fijo. Por tanto, un aumento de la variación dentro  $SSW$  viene acompañada de un descenso en la variación entre  $SSB$ .

Salvo una constante multiplicativa fija,  $SSB$  determina la varianza (4.7), que podemos escribir ahora como

$$\mathbb{V}_{sys}(\hat{Y}_U^{HT}) = N \cdot SSB.$$

En otras palabras, cuanto más homogéneos sean los elementos dentro de las muestras sistemáticas, menos eficiente es el muestreo sistemático. Homogéneo se usa aquí con la connotación de la tendencia a tener iguales los valores de  $y$ . Por tanto, para conseguir una ordenación poblacional favorable para el muestreo sistemático, deberíamos hacer un esfuerzo por conseguir una ordenación que implique un bajo grado de homogeneidad entre los elementos dentro de la misma muestra sistemática.

¿Cómo medimos la homogeneidad? Veamos dos medidas. La primera opción es

$$\rho = 1 - \frac{n}{n-1} \cdot \frac{SSW}{SST}, \quad (4.10)$$

que se denomina *coeficiente de correlación intraclase*, ya que puede escribirse de forma alternativa como

$$\rho = \frac{2 \sum_{r=1}^a \left[ \sum_{k < l} \sum_{s_r} (y_k - \bar{y}_U)(y_l - \bar{y}_U) \right]}{(n-1)(N-1)S_{yU}^2}. \quad (4.11)$$

<sup>3</sup>En inglés, *within systematic samples*.

<sup>4</sup>En inglés, *between systematic samples*.

Podemos interpretar  $\rho$  como una medida de la correlación entre pares de elementos dentro de la misma muestra sistemática. Se obtiene un valor positivo de  $\rho$  cuando los elementos en la misma muestra tienden a tener valores similares de  $y$ . En un extremo,  $\rho = 1$  si  $SSW = 0$ , es decir, hay una homogeneidad completa (no hay variación) dentro de las muestras sistemáticas. En el otro extremo,  $\rho = -\frac{1}{n-1}$  si  $SSB/SCE = 0$ , es decir, heterogeneidad completa dentro de las muestras.

La medida de homogeneidad que preferimos está muy relacionada con  $\rho$ , concretamente, el *coeficiente de homogeneidad* definido por

$$\delta = 1 - \frac{N-1}{N-a} \cdot \frac{SSW}{SST}. \quad (4.12)$$

Introduciendo la *varianza intra-muestral*,

$$S_{yW}^2 = \frac{SSW}{N-a} \quad (4.13)$$

y teniendo en cuenta que la varianza poblacional total es

$$S_{yU}^2 = \frac{SST}{N-1} \quad (4.14)$$

tenemos la ecuación

$$S_{yW}^2 = (1 - \delta) S_{yU}^2$$

o

$$\frac{\text{varianza intramuestral}}{\text{varianza general}} = 1 - \delta. \quad (4.15)$$

Una ventaja con  $\delta$  (en lugar de con  $\rho$ ) como medida de homogeneidad es que la representación de  $\rho$  en la ecuación (4.11), que recuerda al coeficiente de correlación, sólo se verifica para clases (muestras) de igual tamaño. Por contra, la representación de fácil comprensión de  $\delta$  en la ecuación (4.15) se aplica tanto si los  $s_r$  son de igual tamaño como si no. Los valores extremos de  $\delta$  son

$$\delta_{min} = -\frac{a-1}{N-a}$$

que ocurre si  $SSB/SCE = 0$  ( $\bar{y}_s = \text{constante para todo } s$ ), y

$$\delta_{max} = 1$$

que ocurre si  $SSW = 0$ , es decir, hay homogeneidad completa.

El siguiente resultado expresa  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$  como una función de  $\delta$  y el útil para las comparaciones con el muestreo aleatorio simple sin reemplazamiento.

**Teorema 15**

Bajo el diseño muestral sistemático (con  $N = an$ , donde  $a$  es un entero), la varianza del estimador HT del total poblacional se puede escribir como

$$\mathbb{V}_{sys}(\hat{Y}_U^{HT}) = \frac{N^2 S_{yU}^2}{n} [(1 - f) + (n - 1)\delta], \quad (4.16)$$

donde  $f = \frac{n}{N} = \frac{1}{a}$  es la fracción muestral.

**Demostración 15**

La demostración se obtiene usando la ecuación (4.7) y la definición de  $\delta$ .

A partir de (4.16) vemos, una vez más, que cuanto más homogeneidad, o menos heterogeneidad, entre los elementos dentro de las muestras sistemáticas, menos eficiente es el muestreo sistemático.

Comparemos los diseños muestrales sistemático y aleatorio simple sin reemplazamiento. Para acortar la notación, sean  $V_{sys} = \mathbb{V}_{sys}(\hat{Y}_U^{HT})$  y  $V_{srswor} = \mathbb{V}_{srswor}(\hat{Y}_U^{HT})$  las respectivas varianzas del estimador HT del total poblacional  $Y_U = \sum_{k \in U} y_k$ . En ambos casos, el estimador es  $\hat{Y}_U^{HT} = N\bar{y}_s$ . Sin embargo, la distribución en el muestreo es diferente en ambos casos. Puesto que

$$V_{srswor} = N^2 \frac{1-f}{n} S_{yU}^2,$$

el efecto de diseño del muestreo sistemático se obtiene, con la ayuda de (4.16), como

$$\text{deff}_{sys} = \frac{V_{sys}}{V_{srswor}} = 1 + \frac{n-1}{1-f} \delta.$$

Por lo tanto, el muestreo sistemático es más eficiente que el muestreo aleatorio simple sin reemplazamiento si  $\delta < 0$ , es decir, si  $S_{yW}^2 > S_{yU}^2$ . Para crear una situación donde se verifique esta condición, debemos (si es posible) organizar la población de forma que los  $y_k$  dentro de cada muestra sistemática muestre una considerable heterogeneidad. Esto a menudo será el caso si se organiza de forma que los elementos vecinos tengan valores  $y_k$  parecidos entre sí. En cambio, el muestreo aleatorio simple sin reemplazamiento será más eficiente que el muestreo sistemático si  $\delta > 0$ , es decir, si  $S_{yW}^2 < S_{yU}^2$ . Sin embargo, en la práctica, el estadístico a menudo no dispone de la información necesaria para crear una ordenación con las propiedades dadas, por ejemplo, una ordenación favorable el muestreo sistemático. A menudo la ordenación se da de una vez por todas.

*Ejemplo 30.* Este ejemplo muestra los efectos de distintas ordenaciones de la población. Supongamos que  $N = 100$ , que la variable  $y$  toma los valores  $1, 2, \dots, 100$  y que  $n = 10$ .

Entonces el número de muestras es  $a = \frac{N}{n} = 10$ . Tenemos

$$S_{yU}^2 = \frac{N(N+1)}{12} = \frac{100 \cdot 101}{12}.$$

Independientemente de la ordenación de la población, la varianza bajo el muestreo sistemático de tamaño muestral  $n = 10$  a partir de  $N = 100$  es

$$V_{srswor} = N^2 \frac{1-f}{n} S_{yU}^2 = 7,575 \cdot 10^5.$$

Examinemos unas cuantas ordenaciones:

- (a) Supongamos la ordenación en la que  $y_k = k$  ( $k = 1, \dots, 100$ ), es decir, una tendencia lineal perfecta en los valores  $y_k$ . La Tabla 4.2 muestra las diez muestras sistemáticas posibles e  $Y_{sr}$ , que es el total de la muestra  $r$ -ésima.

		$r$									
		1	2	3	4	5	6	7	8	9	10
$y_k$	1	1	2	3	4	5	6	7	8	9	10
	11	11	12	13	14	15	16	17	18	19	20
	21	21	22	23	24	25	26	27	28	29	30
	31	31	32	33	34	35	36	37	38	39	40
	41	41	42	43	44	45	46	47	48	49	50
	51	51	52	53	54	55	56	57	58	59	60
	61	61	62	63	64	65	66	67	68	69	70
	71	71	72	73	74	75	76	77	78	79	80
	81	81	82	83	84	85	86	87	88	89	90
	91	91	92	93	94	95	96	97	98	99	100
$Y_{sr}$		460	470	480	490	500	510	520	530	540	550

Tabla 4.2: Ordenación con tendencia lineal perfecta.

En este caso,  $V_{sys} = 8,25 \cdot 10^4$ , lo que significa que  $V_{srswor}$  es más de nueve veces  $V_{sys}$ . El coeficiente de homogeneidad es  $\delta = -0,089$ , lo que significa que no está lejos de  $\delta_{min} = -0,1$ .

- (b) Una ordenación óptima (varianza mínima) para el muestreo sistemático viene dada en la siguiente Tabla 4.3.

Puesto que todos los totales muestrales  $Y_{sr}$  son iguales,  $V_{sys} = 0$ , y  $\delta = \delta_{min} = -0,1$ .

- (c) Un valor grande positivo del coeficiente de homogeneidad  $\delta$  se asocia con una ordenación como la de la siguiente Tabla 4.4.

		$r$									
		1	2	3	4	5	6	7	8	9	10
$y_k$	1	2	3	4	5	6	7	8	9	10	
	20	19	18	17	16	15	14	13	12	11	
	21	22	23	24	25	26	27	28	29	30	
	40	39	38	37	36	35	34	33	32	31	
	41	42	43	44	45	46	47	48	49	50	
	60	59	58	57	56	55	54	53	52	51	
	61	62	63	64	65	66	67	68	69	70	
	80	79	78	77	76	75	74	73	72	71	
	81	82	83	84	85	86	87	88	89	90	
	100	99	98	97	96	95	94	93	92	91	
$Y_{s_r}$	505	505	505	505	505	505	505	505	505	505	505

Tabla 4.3: Ordenación óptima (varianza mínima).

		$r$									
		1	2	3	4	5	6	7	8	9	10
$y_k$	1	11	21	31	41	51	61	71	81	91	
	2	12	22	32	42	52	62	72	82	92	
	3	13	23	33	43	53	63	73	83	93	
	4	14	24	34	44	54	64	74	84	94	
	5	15	25	35	45	55	65	75	85	95	
	6	16	26	36	46	56	66	76	86	96	
	7	17	27	37	47	57	67	77	87	97	
	8	18	28	38	48	58	68	78	88	98	
	9	19	29	39	49	59	69	79	89	99	
	10	20	30	40	50	60	70	80	90	100	
$Y_{s_r}$	55	155	255	355	455	555	655	755	855	955	

Tabla 4.4: Ordenación con un  $\delta$  grande positivo.

Obtenemos  $V_{sys} = 8,25 \cdot 10^6$ , que es casi 11 veces  $V_{srswor}$ . Aquí  $\delta = 0,989$ , lo que está cerca del máximo  $\delta_{max} = 1$ .

(d) Una ordenación al azar se muestra en la Tabla 4.5.

Aquí,  $V_{sys} = 7,1766 \cdot 10^5$ , que está cerca de  $V_{srswor}$ , y  $\delta = -0,005$ . Esta ordenación fue creada mediante una permutación aleatoria de los enteros de 1 a 100. La población se puede decir que está en orden aleatorio, por lo que, se espera que  $\delta$  esté cerca de 0.

Este ejemplo muestra el cuidado que hay que tener cuando se use el muestreo sistemáti-

	$r$									
	1	2	3	4	5	6	7	8	9	10
$y_k$	48	14	71	13	40	59	18	45	6	53
	38	23	11	58	70	22	24	88	77	84
	10	51	98	65	93	68	25	32	99	9
	17	26	8	78	34	87	96	39	20	54
	56	79	31	86	43	66	2	62	57	5
	73	7	80	27	60	89	76	81	85	83
	3	28	33	90	55	1	21	69	61	92
	74	37	44	94	12	72	100	30	63	97
	75	41	16	82	35	95	67	50	64	29
	49	42	15	19	46	36	47	91	52	4
$Y_{s_r}$	443	348	407	612	488	595	476	587	584	510

Tabla 4.5: Ordenación al azar

co. El precio por la simplicidad del muestreo sistemático puede ser una alta pérdida de eficiencia. Por otro lado, si la población está ordenada de forma favorable al muestreo sistemático, se puede obtener una gran ganancia en la eficiencia.



#### 4.4 Problemática de la estimación de la varianza

Uno de los problemas del muestreo sistemático es que no hay un estimador insesgado de la varianza  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ . Es una contrapartida por la simplicidad del muestreo sistemático. No podemos evaluar la variabilidad muestral de la estimación puntual. Existen algunas aproximaciones que no son perfectas para tratar este problema, una de ellas es usar un estimador sesgado de la varianza; otro es modificar la selección sistemática para permitir una estimación insesgada de la varianza.

Se han propuesto varios estimadores sesgados de la varianza; mencionaremos uno de ellos. Supongamos que hay una razón de peso para creer que, en una aplicación específica, el muestreo sistemático es por lo menos tan bueno como el muestreo aleatorio simple sin reemplazamiento cuando la varianza pequeña es el criterio. Si  $s_r$  es la muestra sistemática seleccionada, la varianza muestral es

$$S_{y_{s_r}}^2 = \frac{1}{n-1} \sum_{s_r} (y_k - \bar{y}_{s_r})^2.$$

Como estimador de  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$  podemos considerar

$$\hat{V} = \frac{N^2(1-f)}{n} S_{y_{s_r}}^2 \quad (4.17)$$



que es el estimador de la varianza apropiado para el diseño aleatorio simple sin reemplazamiento.

Supongamos que estamos en una situación en la que el muestreo sistemático es más eficiente que el muestreo aleatorio simple sin reemplazamiento, es decir,

$$\mathbb{V}_{sys}(\hat{Y}_U^{HT}) < \mathbb{V}_{srswor}(\hat{Y}_U^{HT}),$$

lo que se verifica si y solo si  $\delta < 0$ . Entonces se puede demostrar que  $\hat{V}$ , dada por la ecuación (4.17), *sobreestimar*á  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ . Un estimador de la varianza  $\hat{V}$  se dice que *sobreestima* si su esperanza es superior a la varianza para la cual  $\hat{V}$  se usa como estimador. Es decir, para el estimador  $\hat{V}$  dado por la ecuación (4.17) tenemos

$$\mathbb{E}_{sys}(\hat{V}) > \mathbb{V}_{sys}(\hat{Y}_U^{HT}),$$

si  $\delta < 0$ . Cuando se usa el muestreo sistemático y  $\delta < 0$ , un intervalo de confianza calculado para  $Y_U$  mediante la ecuación (4.17) será (lo que se califica como) *conservador*. Un intervalo de confianza se denomina conservador si al considerar un nivel de confianza de, por ejemplo, 95 % para el cálculo del intervalo (usaríamos entonces la constante  $z_{0,975} = 1,96$ , suponiendo normalidad aproximada) y el nivel de confianza real es mayor que 95 %. En muestras repetidas, este intervalo contiene el valor de un parámetro desconocido a una tasa superior al 95 % de todas las muestras.

*Ejemplo 31.* Para la población en el Ejemplo 30 tenemos (aproximadamente) la Tabla 30:

Orden	$\mathbb{E}_{sys}(\hat{V})$	$\mathbb{V}_{sys}(\hat{Y}_U^{HT})$	
(a)	$8,3 \cdot 10^6$	$8,3 \cdot 10^4$	sobreestimación
(b)	$8,3 \cdot 10^5$	0	sobreestimación
(c)	$8,3 \cdot 10^3$	$8,3 \cdot 10^6$	subestimación
(d)	$8,3 \cdot 10^5$	$7,2 \cdot 10^5$	aproximadamente insesgado

Tabla 4.6: Valores esperados de  $\hat{V}$  para las muestras sistemáticas del Ejemplo 30.

En el caso (d), en que el muestreo sistemático y el muestreo aleatorio simple sin reemplazamiento son diseños aproximadamente igual de eficientes (y  $\delta$  muy cerca de cero), hay concordancia entre  $\mathbb{E}_{sys}(\hat{V})$  y  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ . En los otros casos,  $\hat{V}$  o bien sobreestima o bien subestima  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ , dependiendo de si  $\delta < 0$  o  $\delta > 0$ . Esto se ve confirmado por los datos de la tabla. Por ejemplo, en el caso (a), el valor de  $\mathbb{E}_{sys}(\hat{V})$  es mucho mayor que el de  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ , mientras que la relación es inversa en el caso (c). Se han propuesto alternativas a  $(\hat{V})$ . En [Wolter 2007](#) hay un capítulo entero sobre la estimación de la varianza del muestreo sistemático. ■

**Comentario 12.** El muestreo sistemático es más preciso que el muestreo aleatorio simple cuando la varianza dentro de las posible muestras sistemáticas es *mayor* que la varianza

general de la población (en este caso, las medias de las posibles muestras sistemáticas son más similares). Si existe poca variación dentro de las muestras sistemáticas con respecto a la población correspondiente, entonces todos los elementos de la muestra dan una información similar y es de esperar que el muestreo sistemático tenga una varianza mayor que una muestra aleatoria simple. ■

**Comentario 13.** Las 4 ordenaciones de los ejercicios previos se pueden resumir en tres estructuras de población distintas:

1. *La población tiene un orden aleatorio.* Es probable que el muestreo sistemático produzca una muestra que se comporta como una muestra aleatoria simple. En muchos casos, el orden de la población no está relacionado con las características de interés, como cuando en el marco de personas ésta figura en orden alfabético. En un caso como éste, no hay razón para creer que las personas seleccionadas para una muestra sistemática serán más o menos similares a las de una muestra aleatoria de personas. En este caso, el muestreo aleatorio simple y el sistemático darán resultados similares. Se podrían utilizar los resultados y las fórmulas del muestreo aleatorio simple para estimar  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$ .
2. *El marco muestral tiene un orden creciente o decreciente.* Es probable que el muestreo sistemático sea más preciso que el muestreo aleatorio simple. Algunos registros contables pueden estar enumerados con las cantidades más pequeñas al principio y las grandes al final o viceversa. En tal caso decimos que la población tiene una autocorrelación negativa o positiva, respectivamente: los elementos adyacentes tienden a ser más similares que los elementos más alejados. En este caso,  $\mathbb{V}_{sys}(\hat{Y}_U^{HT})$  es menor que la varianza de la media muestral en una muestra aleatoria simple del mismo tamaño. Una muestra sistemática obliga a que los valores de la muestra se dispersen; es posible que una muestra aleatoria simple seleccione todas las unidades muestrales de entre valores mayores o de entre los menores. Cuando el marco muestral tiene un orden creciente o decreciente no es conveniente usar las fórmulas del muestreo aleatorio simple, como en el caso 1, ya que se puede producir una sobreestimación y los intervalos de confianza así contruidos pueden no ser acurados.

En un caso como éste el muestreo estratificado <sup>5</sup> puede funcionar mejor que el muestreo sistemático para las poblaciones con autocorrelación negativa o positiva: si el arranque aleatorio está cerca de cualquier de los dos extremos del intervalo de muestreo, una muestra sistemática tenderá a dar una estimación demasiado pequeña o demasiado grande.

3. *El marco muestral tiene un patrón periódico.* Si extraemos una muestra con el mismo intervalo que la periodicidad, el muestreo sistemático será menos preciso que el muestreo aleatorio simple. El muestreo sistemático no es tan acurado cuando la

<sup>5</sup>El muestreo estratificado se ve en el Tema 6 de este bloque.

población tiene un orden cíclico o periódico y el intervalo de muestreo coincide con un múltiplo del periodo.

Supongamos que los valores de la población (en orden) son:

1 2 3 1 2 3 1 1 3 1 2 3

y que el intervalo de muestreo es 3. Entonces, todos los elementos de la muestra sistemática serán iguales; si usamos la fórmula de la muestra aleatoria simple para estimar la varianza, tendremos que  $\widehat{V}_{sys}(\widehat{Y}_U^{HT}) = 0$ . Pero el valor real de  $V_{sys}(\widehat{Y}_U^{HT})$  para esta población es  $\frac{2}{3}$ . Por tanto, la muestra sistemática no sería más precisa que una única observación elegida al azar entre la población.

■

Otra solución al problema de la estimación de la varianza es modificar el diseño muestral sistemático. Por ejemplo, en lugar de usar solo un arranque aleatorio y el intervalo muestral  $a$ , podemos usar  $m > 1$  arranques aleatorios y el intervalo muestral  $ma$ . Esto nos proporciona una muestra que consiste en  $m$  'porciones' sistemáticas, cada una de un tamaño  $\frac{n}{m}$ . Asumamos por simplicidad que  $\frac{n}{m}$  y  $a = \frac{N}{n}$  son enteros. Una muestra sistemática de  $m$  enteros se selecciona de los enteros 1 a  $ma$ , incluido. Digamos que los números seleccionados son  $r_1, r_2, \dots, r_m$ . La muestra entonces viene dada por

$$s = \{k : k = r_i + (j - 1)ma; i = 1, \dots, m, j = 1, \dots, \frac{n}{m}\}.$$

En este caso  $\pi_k = \frac{m}{ma} = \frac{n}{N}$  para cada  $k$ , mientras que

$$\pi_{kl} = \begin{cases} \frac{n}{N} & \text{si } k \text{ y } l \text{ pertenecen a la misma muestra } s, \\ \frac{n}{N} \cdot \frac{m-1}{ma-1} & \text{si } k \text{ y } l \text{ pertenecen a muestras distintas.} \end{cases}$$

En la terminología del muestreo por conglomerados monoetápico (véase el Tema 7 de este mismo bloque), el muestreo sistemático con  $m$  arranques aleatorios es equivalente a agrupar la población en  $ma$  conglomerados, cada uno de tamaño  $\frac{n}{m}$  y tomar una muestra aleatoria simple sin reemplazamiento de  $m$  conglomerados. Todos los elementos de los conglomerados seleccionados son encuestados.

La variedad de arranques aleatorios unitarios del muestreo sistemático es también un caso especial de muestreo por conglomerados monoetápico, concretamente, con un único conglomerado seleccionado aleatoriamente. Un inconveniente que nos encontramos con varios arranques aleatorios es que a menudo da lugar a una varianza mayor que un único arranque aleatorio.

Como se puede ver en el muestreo por conglomerados monoetápico y en el muestreo bi- y multi-etápico, en algunas encuestas, el problema de la estimación de la varianza no es tan serio como pueda parecer. A menudo el muestreo sistemático se usa para la selección de las unidades de última etapa en un diseño muestral multietápico. En tales situaciones, a menudo es posible obtener buenos estimadores de la varianza para el estimador HT a pesar del uso de muestreo sistemático en la última etapa.

El muestreo sistemático en dos dimensiones es apropiado en algunas aplicaciones. Un campo o un bosque pueden estar dividido en pequeñas cuadrículas o en rectángulos de igual tamaño. Un arranque aleatorio puntual (que consiste en una coordenada horizontal aleatoria y una coordenada vertical aleatoria) se determina para la primera cuadrícula; el punto seleccionado en la primera cuadrícula entonces identificará sistemáticamente un punto muestral en cada una de las otras cuadrículas. [Bellhouse 1977](#) (véase también [Bellhouse 1981](#)) desarrolló diseños para muestreo espacial; un resumen del muestreo sistemático en una o más dimensiones se pueden encontrar en ([Bellhouse 1981](#)).

**Comentario 14. Relación entre el muestreo sistemático y el muestreo por conglomerados**<sup>6</sup>. Si la periodicidad es una preocupación, una solución consiste en usar las **muestras sistemáticas interpenetrantes** ([Mahalanobis 1946](#)). En vez de tomar una sola muestra sistemática de la población se toman varias. Entonces se podrán usar las fórmulas para las muestras por conglomerado a las distintas muestras sistemáticas y estimar las varianzas; cada muestra sistemática se comportará como un conglomerado. ■

## Bibliografía

- Bellhouse, D.R. (1977). "Some optimal designs for sampling in two dimensions". En: *Biometrika* 64, págs. 605-611.
- (1981). "Spatial sampling in the presence of a trend". En: *Journal of Statistical Planning and Inference* 5, págs. 365-375.
- Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.
- Mahalanobis, P.C. (1946). "Recent experiment in statistical sampling in the Indian Statistical Institute". En: *Jours. Roy. Stat.Soc.* 109, págs. 325-370.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
- Wolter, K. (2007). *Introduction to variance estimation*. 2nd. New York: Springer.

---

<sup>6</sup>El muestreo por conglomerados se ve en el Tema 7 de este mismo bloque

## Tema 5

### Muestreo de Poisson: definición, estimadores, varianza y estimador de la varianza. Muestreo con probabilidades proporcionales al tamaño: muestreo sin reemplazamiento y con reemplazamiento

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

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

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.

#### 5.1 Introducción

El muestreo de Poisson y el muestreo con probabilidades proporcionales al tamaño son dos tipos de diseños de muestreo directo de elementos. Como ya se comentó en el Tema 3, los diseños de muestreo directo de elementos presentan dos características importantes:

1. Existe un marco de muestreo que identifica a cada elemento de la población.
2. Las unidades de muestreo son precisamente los elementos de la población.

Tanto el muestreo de Poisson como el muestreo con probabilidades proporcionales al tamaño pertenecen a la categoría de diseños denominados *diseños de muestreo con probabilidades desiguales*. La característica común de este tipo de diseños es que las probabilidades de inclusión de primer orden  $\pi_k$  son distintas, a diferencia de los *diseños de muestreo con probabilidades iguales*, donde  $\pi_k = \text{constante}$ , para todo  $k = 1, \dots, N$ .

## 5.2 Muestreo de Poisson

### 5.2.1 Definición

El muestreo de Poisson es un diseño de muestreo con probabilidades desiguales que produce muestras de tamaño variable. Se trata de una generalización del muestreo de Bernoulli. Denotamos el muestreo de Poisson como diseño *Prs*.

Sea  $\pi_k = \mathbb{P}(u_k \in s)$  la probabilidad de inclusión del elemento  $k$ . Bajo un muestreo de Poisson los indicadores de pertenencia a la muestra  $I_k(s) = I_k$  son variables aleatorias independientes con distribución

$$\mathbb{P}(I_k = 1) = \pi_k, \quad \mathbb{P}(I_k = 0) = 1 - \pi_k, \quad \text{para todo } k = 1, \dots, N.$$

Las probabilidades de inclusión de segundo orden son, como consecuencia de la independencia,

$$\pi_{kl} = \mathbb{E}(I_k \cdot I_l) = \mathbb{E}(I_k) \cdot \mathbb{E}(I_l) = \pi_k \cdot \pi_l, \quad \text{para todo } k \neq l.$$

Cualquier posible muestra es una secuencia de  $N$  experimentos independientes. La probabilidad de seleccionar una muestra es  $\prod_{k \in s} \pi_k \prod_{k \in U-s} (1 - \pi_k)$ . Por tanto, el diseño muestral de Poisson es

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U-s} (1 - \pi_k), \quad \text{para todo } s \in \Omega,$$

donde  $\Omega$  es el espacio muestral, el conjunto de posibles muestras.

Sea  $n_s$  la variable aleatoria que representa el número de elementos seleccionados en la muestra aleatoria  $s$ , cuya media y varianza son, respectivamente,

$$\begin{aligned} \mathbb{E}(n_s) &= \sum_{k \in U} \pi_k \\ \mathbb{V}(n_s) &= \sum_{k \in U} \pi_k \cdot (1 - \pi_k) \end{aligned}$$

*Ejemplo 32.* Un ejemplo de implementación del diseño de *Prs* es el siguiente:

- Se consideran  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo  $(0, 1) : \varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ .
- Dado un conjunto de probabilidades de inclusión  $\pi_1, \pi_2, \dots, \pi_N$ , si  $\varepsilon_k < \pi_k$ , entonces el elemento  $u_k$  es seleccionado en la muestra, en caso contrario no.

■

### 5.2.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador de Horvitz-Thompson (HT) como estimador del total poblacional  $Y_U$ . Veamos a continuación cuál es el estimador y la varianza bajo el diseño  $Prs$ . Para demostrar los resultados se hará uso de las expresiones obtenidas en el Tema 2.

#### Proposición 16

Bajo el diseño de Poisson, el estimador HT de  $Y_U$  viene dado por

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} \quad (5.1)$$

La varianza del estimador  $\hat{Y}_U^{\text{HT}}$  es

$$\mathbb{V}_{Prs}(\hat{Y}_U^{\text{HT}}) = \sum_{k \in U} \left( \frac{1}{\pi_k} - 1 \right) y_k^2$$

Un estimador insesgado de la varianza del estimador es

$$\hat{\mathbb{V}}_{Prs}(\hat{Y}_U^{\text{HT}}) = \sum_{k \in s} \frac{1}{\pi_k} \cdot \left( \frac{1}{\pi_k} - 1 \right) y_k^2 \quad (5.2)$$

#### Demostración 16

El estimador de Horvitz-Thompson es, por definición, el que se muestra en (5.1).

La varianza del estimador es

$$\begin{aligned} \mathbb{V}_{Prs}(\hat{Y}_U^{\text{HT}}) &= \sum_{k \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \cdot (1 - \pi_k) + \underbrace{\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l)}_0 = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k} \cdot (1 - \pi_k) = \sum_{k \in U} \left( \frac{1}{\pi_k} - 1 \right) y_k^2 \end{aligned}$$

Por último, se demuestra la insesgader del estimador (5.2):

$$\begin{aligned} \mathbb{E}[\hat{\mathbb{V}}_{Prs}(\hat{Y}_U^{\text{HT}})] &= \mathbb{E}\left[\sum_{k \in s} \frac{1}{\pi_k} \cdot \left( \frac{1}{\pi_k} - 1 \right) \cdot y_k^2\right] = \\ &= \sum_{k \in U} \frac{1}{\pi_k} \cdot \left( \frac{1}{\pi_k} - 1 \right) \cdot y_k^2 \cdot \mathbb{E}(I_k) = \end{aligned}$$

$$= \sum_{k \in U} \left( \frac{1}{\pi_k} - 1 \right) y_k^2$$

Como es de esperar, bajo el diseño de Poisson la varianza del estimador HT podría ser excesivamente grande debido a la variabilidad del tamaño de la muestra. Una alternativa es usar el estimador mejorado siguiente, ya que aunque es ligeramente sesgado, normalmente se obtiene una varianza menor:

$$\hat{Y}_U^{Rat} = N \cdot \frac{\sum_{k \in s} \frac{y_k}{\pi_k}}{\hat{N}}$$

donde  $\hat{N} = \sum_{k \in s} \frac{1}{\pi_k}$ .

**Comentario 15.** Se puede demostrar que la varianza aproximada  $AV$  del estimador  $\hat{Y}_U^{Rat}$  viene dada por

$$AV \left( \hat{Y}_U^{Rat} \right) = \sum_{k \in U} \frac{(y_k - \bar{y}_U)^2}{\pi_k} - N \cdot S_{yU}^2$$

Véase el Tema 3<sup>1</sup> del bloque “Producción Estadística Oficial: Métodos Avanzados” del grupo de materias específicas, sobre la estimación de razón. ■

Volvamos de nuevo al estimador HT, nos preguntamos ahora qué valor deben tomar las probabilidades de inclusión  $\pi_k$  cuando el tamaño de muestra esperado es fijo,  $n$ . Podrían tomarse por ejemplo aquellos valores que minimicen la varianza del estimador sujeto a la restricción  $n = \sum_{k \in s} \pi_k$ . Esto es:

$$\begin{cases} \text{minimizar } \mathbb{V}_{Prs} \left( \hat{Y}_U^{HT} \right) = \sum_{k \in U} \left( \frac{1}{\pi_k} - 1 \right) y_k^2 \\ \text{sujeto a } n = \sum_{k \in s} \pi_k \end{cases} \quad (5.3)$$

El problema es equivalente a minimizar el producto

$$\left( \sum_{k \in U} \frac{y_k^2}{\pi_k} \right) \cdot \left( \sum_{k \in U} \pi_k \right) \geq \left( \sum_{k \in U} y_k \right)^2$$

De la desigualdad de Cauchy se tiene

$$\left( \sum a_k^2 \right) \cdot \left( \sum b_k^2 \right) \geq \left( \sum a_k \cdot b_k \right)^2.$$

<sup>1</sup>Tema 3. Introducción a problemas de estimación complejos. El efecto del sesgo en intervalos de confianza de las estimaciones. Consistencia e insesgadez asintótica. La técnica de linealización de Taylor para la estimación de la varianza. Estimador de una razón: varianza y sesgo.



La igualdad se da si y solo si  $\frac{b_k}{a_k}$  es constante para cada  $k$ . Tomando  $a_k = \frac{y_k^2}{\pi_k}$  y  $b_k = \pi_k$ , se da la igualdad si y solo si  $\frac{y_k}{\pi_k} = \lambda$  es constante. Suponiendo que  $y_k > 0$  para todo  $k$ , se tiene  $\pi_k = \frac{y_k}{\lambda}$ . Dado que  $n = \sum_{k \in U} \pi_k$ , se llega a

$$\pi_k = n \cdot \frac{y_k}{\sum_{k \in U} y_k},$$

suponiendo  $y_k \leq \sum_{k \in U} \frac{y_k}{n}$ ,  $\forall k$ . Si se eligen  $\pi_k$  de esta forma, el estimador HT queda

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{n_s}{n} \sum_{k \in U} y_k$$

con  $n = \sum_{k \in U} \pi_k$  el tamaño de muestra esperado. Por tanto, la variación del estimador HT dependerá simplemente de la variación en el tamaño de la muestra  $n_s$ .

**Comentario 16.** Bajo un diseño de muestreo que produce muestras de tamaño fijo, si  $\pi_k$  es exactamente proporcional a  $y_k$ , el estimador HT tendría varianza cero. ■

Dado que las probabilidades de inclusión obtenidas dependen de  $y_k$ , generalmente no podrá utilizarse esto en la práctica. Sin embargo, si se dispone de una variable auxiliar  $x$  cuyos valores son positivos y conocidos para toda la población, y además se sospecha que  $y$  es aproximadamente proporcional a  $x$ , entonces podría tomarse  $\pi_k$  de la siguiente forma

$$\pi_k = n \cdot \frac{x_k}{\sum_{k \in U} x_k} \quad (5.4)$$

suponiendo  $x_k \leq \sum_{k \in U} \frac{x_k}{n}$ ,  $\forall k$ . En caso de no serlo, se tomaría  $\pi_k = 1$ . En este caso, el estimador HT tendrá una varianza pequeña.

Las probabilidades de inclusión determinadas según (5.4) se dice que son *proporcionales al tamaño*, ya que  $x_k$  se considera una medida del tamaño del elemento  $k$ . A continuación se explicará en detalle el diseño de muestreo con probabilidades proporcionales al tamaño.

### 5.3 Muestreo con probabilidades proporcionales al tamaño

El muestreo con probabilidades proporcionales al tamaño es un *diseño de muestreo con probabilidades desiguales* donde las probabilidades de inclusión de cada elemento de la población son proporcionales al valor que toma dicho elemento en una variable  $x$ , esto es,

$$\pi_k \propto x_k$$

donde  $x_k$  es un valor positivo, para todo  $k = 1, \dots, N$ . El valor de la variable auxiliar  $x$  debe ser conocido para todos los elementos de la población y debe ser elegida de forma que sea aproximadamente proporcional a la variable de estudio  $y$ .

En el caso de un diseño con reemplazamiento, se tiene que las probabilidades de selección de cada elemento de la población son proporcionales al valor que toma dicho elemento en la variable  $x$ , esto es,

$$p_k \propto x_k, \text{ para valores positivos y conocidos.}$$

La variable  $x$  se considera una medida del tamaño del elemento  $k$ . Por ejemplo, supongamos que se desean seleccionar secciones censales, podríamos tomar como variable auxiliar el número de viviendas familiares principales por sección censal, como en el caso de la Encuesta de Población Activa<sup>2</sup>.

Estudiaremos el muestreo con probabilidades proporcionales al tamaño para diseños que producen muestras de tamaño fijo  $n$ . El caso sin reemplazamiento se denotará como diseño *ppswor* y el caso con reemplazamiento diseño *ppswr*.

### 5.3.1 Muestreo sin reemplazamiento

En primer lugar se mostrará el incentivo de usar diseños tales que  $\pi_K \propto x_k$  a través de la siguiente proposición.

#### Proposición 17

Bajo un diseño muestral con tamaño de muestra fijo  $n$ , la varianza del estimador de Horvitz-Thompson cuando  $y_k$  es proporcional a  $\pi_k$ , esto es,  $y_k = c \cdot \pi_k$ , es cero.

#### Demostración 17

Dado que se están considerando diseños con un tamaño de muestra fijo, podemos escribir la expresión de la varianza del estimador HT de la forma alternativa dada en (2.5):

$$\mathbb{V}(\hat{Y}_U^{\text{HT}}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.$$

Sustituyendo  $y_k = \pi_k \cdot c$ , se tiene que la varianza es cero.

<sup>2</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176918&menu=metodologia&idp=1254735976595](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=metodologia&idp=1254735976595)

De forma alternativa, podemos ver que el estimador HT es

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{c \cdot \pi_k}{\pi_k} = n \cdot c$$

Por tanto, la varianza del estimador es cero.

Un diseño tal que  $\pi_k \propto y_k$  no puede ser encontrado en la práctica, ya que habitualmente los valores  $y_k$  son desconocidos. En su lugar, se tomará  $\pi_k \propto x_k$  como ya se ha comentado, ya que si  $\pi_k$  es aproximadamente proporcional a  $x_k$ , esto conducirá a cocientes  $\frac{y_k}{\pi_k}$  aproximadamente constantes, de forma que la varianza del estimador HT presentará un valor pequeño.

La condición  $\pi_k \propto x_k$  es equivalente a decir

$$\pi_k = n \cdot \frac{x_k}{\sum_{k \in U} x_k}$$

ya que  $\sum_{k \in U} \pi_k = n$ .

**Comentario 17.** En la condición anterior se debe suponer que  $x_k \leq \sum_{k \in U} \frac{x_k}{n}$ , ya que en caso contrario no se verificaría  $\pi_k \leq 1$ . Si existe un subconjunto de elementos de la población para los que  $x_k > \sum_{k \in U} \frac{x_k}{n}$ , entonces no será posible seleccionar la muestra con  $\pi_k$  estrictamente proporcional a  $x_k$ . En esta situación podemos actuar de la siguiente forma:

- Sea  $A$  el conjunto de elementos tales que  $n \cdot x_k > \sum_{k \in U} x_k$ .
- Tomar

$$\begin{aligned} \pi_k &= 1, \quad \forall k \in A \\ \pi_k &= (n - n_A) \cdot \frac{x_k}{\sum_{k \notin A} x_k}, \quad \forall k \notin A \end{aligned}$$

- Si todos los elementos verifican  $\pi_k \leq 1$ , hemos acabado. En caso contrario, se repite el proceso descrito hasta que se cumpla la condición.

■

A continuación se discutirá el problema de cómo encontrar un mecanismo de selección de la muestra sin reemplazamiento de forma que se tomen las probabilidades de inclusión  $\pi_k$  proporcionales a  $x_k$ . Veremos que no es fácil encontrar un esquema de implementación sencillo. Antes de ello, se enumeran las propiedades deseables para un mecanismo de selección de la muestra:

1. La selección de la muestra es un proceso relativamente sencillo.

2. Las probabilidades de inclusión de primer orden  $\pi_k$  son estrictamente proporcionales a  $x_k$ .
3. Las probabilidades de inclusión de segundo orden satisfacen  $\pi_{kl} > 0$  para todo  $k \neq l$ , ya que en caso de no serlo, no será posible construir una estimación insesgada de la varianza.
4. Las probabilidades de inclusión de segundo orden pueden ser obtenidas sin la necesidad de muchos cálculos pesados.
5.  $\pi_{kl} - \pi_k \cdot \pi_l < 0$  para todo  $k \neq l$  para garantizar que el estimador de la varianza debido a [Yates y Grundy 1953](#) y [Sen 1953](#) tome siempre un valor no negativo, ya que

$$\widehat{V}(\widehat{Y}_U^{\text{HT}}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (5.5)$$

### Método de selección para tamaño de muestra $n = 1$

Sea  $T_k$  el tamaño acumulado hasta el elemento  $k$ , donde  $T_0 = 0$ .

1. Se calcula  $T_k = T_{k-1} + x_k$
2. Se considera una realización de una uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon$ . El elemento  $k$  es seleccionado si

$$T_{k-1} < \varepsilon \cdot T_N \leq T_k$$

Se puede comprobar que se tiene efectivamente un diseño *ppswor*:

$$\pi_k = \mathbb{P}(T_{k-1} < \varepsilon \cdot T_N \leq T_k) = \frac{T_k - T_{k-1}}{T_N} = \frac{x_k}{\sum_{k \in U} x_k}$$

donde  $T_N$  es la suma de todos los valores de la variable auxiliar  $x$ :  $T_N = \sum_{k \in U} x_k$ .

Denotaremos a este esquema como el *método del total acumulado*.

*Ejemplo 33.* Supóngase que una población está formada por 6 elementos para la cual se dispone de la información de una variable auxiliar  $x$  que toma los valores  $x_1 = 15, x_2 = 10, x_3 = 30, x_4 = 20, x_5 = 15$  y  $x_6 = 10$ . Calculemos el total acumulado  $T_k$ :

$k$	$x_k$	$T_k$
1	15	15
2	10	25
3	30	55
4	20	75
5	15	90
6	10	100

Se obtiene una realización de una uniforme sobre el intervalo  $(0, 1)$ ,  $\varepsilon$ . Suponiendo que  $25 < \varepsilon \cdot 100 \leq 55$ , entonces el elemento seleccionado es  $k = 3$ .

■

**Comentario 18.** Obsérvese que las probabilidades de inclusión de segundo orden  $\pi_{kl}$  son cero, para todo  $k \neq l$ , por lo que no es posible dar una estimación insesgada de la varianza. Una condición necesaria pero no suficiente para que  $\pi_{kl} > 0$  es que el tamaño de muestra sea al menos dos.

■

### Método de selección para tamaño de muestra $n = 2$

Se presentará el mecanismo de selección propuesto por [Brewer 1963](#) y [Brewer 1975](#), que verifica todas las propiedades que se enumeraron más arriba para un esquema de selección deseable.

1. La primera unidad, denotada por  $k_1$ , se extrae con probabilidad proporcional a  $c_k$ , donde

$$c_k = \frac{x_k \cdot (T_N - x_k)}{T_N \cdot (T_N - 2 \cdot x_k)}$$

2. La segunda unidad se extrae sin reemplazamiento y con probabilidad

$$p_{l|k_1} = \frac{x_l}{T_N - x_{k_1}}$$

### Proposición 18

En el esquema de Brewer, las probabilidades de inclusión de primer orden  $\pi_k$  verifican que son estrictamente proporcionales a  $x_k$ :

$$\pi_k = 2 \cdot \frac{x_k}{T_N}$$

Las probabilidades de inclusión de segundo orden  $\pi_{kl}$  son

$$\pi_{kl} = \frac{2 \cdot x_k \cdot x_l}{T_N \sum_{k \in U} c_k} \cdot \frac{T_N - x_k - x_l}{(T_N - 2 \cdot x_k) \cdot (T_N - 2 \cdot x_l)}, \quad k \neq l.$$

**Demostración 18**

Dado que la primera unidad se selecciona con probabilidad proporcional a  $c_k$ , esto es,  $p_k \propto c_k$ , entonces la probabilidad de que el elemento  $k$  sea seleccionado es

$$p_k = \frac{c_k}{\sum_{k \in U} c_k}$$

$$\begin{aligned} \pi_k &= \mathbb{P}(u_k \text{ seleccionado en 1ª extracción}) + \\ &+ \sum_{l \neq k} \mathbb{P}(u_l \text{ seleccionado en 1ª extracción y } u_k \text{ en 2ª extracción}) = \\ &= p_k + \sum_{l \neq k} p_l \cdot p_{k|l} = \frac{c_k}{\sum_{k \in U} c_k} + \sum_{l \neq k} \frac{c_l}{\sum_{k \in U} c_k} \cdot \frac{x_k}{T_N - x_l} = \\ &= \frac{1}{\sum_{k \in U} c_k} \cdot \left[ c_k + x_k \sum_{l \neq k} \frac{c_l}{T_N - x_l} \right] = \\ &= \frac{1}{\sum_{k \in U} c_k} \cdot \left[ \frac{x_k \cdot (T_N - x_k)}{T_N \cdot (T_N - 2 \cdot x_k)} + x_k \sum_{l \neq k} \frac{x_l}{T_N \cdot (T_N - 2 \cdot x_l)} \right] = \\ &= \frac{x_k/T_N}{\sum_{k \in U} c_k} \cdot \left[ \frac{T_N - 2 \cdot x_k + x_k}{T_N - 2 \cdot x_k} + \sum_{l \neq k} \frac{x_l}{T_N - 2 \cdot x_l} \right] = \\ &= \frac{x_k/T_N}{\sum_{k \in U} c_k} \cdot \left[ 1 + \sum_{l \in U} \frac{x_l}{T_N - 2 \cdot x_l} \right] = 2 \cdot \frac{x_k}{T_N} \end{aligned}$$

ya que

$$\begin{aligned} \sum_{k \in U} c_k &= \sum_{k \in U} \frac{x_k \cdot (T_N - x_k)}{T_N \cdot (T_N - 2 \cdot x_k)} = \frac{1}{2} \sum_{k \in U} \frac{x_k}{T_N} \cdot \frac{T_N + T_N - 2 \cdot x_k}{T_N - 2 \cdot x_k} = \\ &= \frac{1}{2} \sum_{k \in U} \frac{x_k}{T_N} \cdot \left[ 1 + \frac{x_k}{T_N - 2 \cdot x_k} \right] = \frac{1}{2} \cdot \left[ 1 + \sum_{k \in U} \frac{x_k}{T_N - 2 \cdot x_k} \right] \end{aligned}$$

Por otra parte, las probabilidades de inclusión de segundo orden son

$$\begin{aligned} \pi_{kl} &= p_k \cdot p_{l|k} + p_l \cdot p_{k|l} = \frac{c_k}{\sum_{k \in U} c_k} \cdot \frac{x_l}{T_N - x_k} + \frac{c_l}{\sum_{l \in U} c_l} \cdot \frac{x_k}{T_N - x_l} = \\ &= \frac{1}{\sum_{k \in U} c_k} \cdot \left[ \frac{x_k \cdot x_l}{T_N \cdot (T_N - 2 \cdot x_k)} + \frac{x_k \cdot x_l}{T_N \cdot (T_N - 2 \cdot x_l)} \right] = \\ &= \frac{2 \cdot x_k \cdot x_l}{T_N \sum_{k \in U} c_k} \cdot \left[ \frac{T_N - x_k - x_l}{(T_N - 2 \cdot x_k) \cdot (T_N - 2 \cdot x_l)} \right] \end{aligned}$$

**Comentario 19.** Este esquema satisface  $\pi_{kl} - \pi_k \cdot \pi_l < 0$  para todo  $k \neq l$  (véase Rao 1965). Por tanto, el estimador de la varianza dado en (5.5) es siempre no negativo. ■

*Ejemplo 34.* Se dispone de una población constituida por 284 municipios (Apéndice B de Särndal, Swensson y Wretman 1992). Considérense los cuatro municipios de mayor población. Se desea estimar el total de individuos en el año 1985. Para ello se aplica un muestreo con probabilidades proporcionales al tamaño, donde la variable auxiliar usada es la población en 1975 en cada uno de los municipios ( $x$ ), cuyos valores son  $x_1 = 671, x_2 = 446, x_3 = 247, x_4 = 138$ . Se extrae una muestra de tamaño dos usando el esquema de Brewer. Los resultados de la muestra se resumen en la siguiente tabla:

$k$	$y_k$	$x_k$
1	653	671
3	229	247

Se han seleccionado los municipios 1 y 3. Una estimación insesgada del total de la población en 1985 es

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{2x_k/T_N} = \frac{653}{1342/1502} + \frac{229}{494/1502} \approx 1427,125$$

Para obtener un estimador insesgado de la varianza debemos calcular en primer lugar la probabilidad de inclusión de segundo orden de los elementos 1 y 3:

$$\pi_{13} = \frac{2 \cdot x_k \cdot x_l}{T_N \sum_{k \in U} c_k} \cdot \left[ \frac{T_N - x_k - x_l}{(T_N - 2 \cdot x_k) \cdot (T_N - 2 \cdot x_l)} \right]$$

$k$	$x_k$	$T_k$	$c_k$
1	671	671	$\frac{671 \cdot (1502 - 671)}{1502 \cdot (1502 - 2 \cdot 671)}$
2	446	1117	$\frac{446 \cdot (1502 - 446)}{1502 \cdot (1502 - 2 \cdot 446)}$
3	247	1364	$\frac{247 \cdot (1502 - 247)}{1502 \cdot (1502 - 2 \cdot 247)}$
4	138	1502	$\frac{138 \cdot (1502 - 138)}{1502 \cdot (1502 - 2 \cdot 138)}$

Por tanto, se tiene:

$$\pi_{13} = \frac{2 \cdot 671 \cdot 247}{1502 \sum_{k \in U} c_k} \cdot \frac{1502 - 671 - 247}{(1502 - 1342) \cdot (1502 - 494)} \approx 0,254$$

donde  $\sum_{k \in U} c_k \approx 3,141$ .

$$\begin{aligned}\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}}) &= -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 = -\frac{\pi_{13} - \pi_1 \cdot \pi_3}{\pi_{13}} \cdot \left( \frac{y_1}{\pi_1} - \frac{y_3}{\pi_3} \right)^2 \approx \\ &\approx -\left( 1 - \frac{(1342/1502) \cdot (494/1502)}{0,254} \right) \cdot \left( \frac{653}{1342/1502} - \frac{229}{494/1502} \right)^2 \approx 185,527\end{aligned}$$

■

### Método de selección para tamaño de muestra $n > 2$

La mayoría de los algoritmos de selección de la muestra que implementan el diseño de muestreo con probabilidades proporcionales al tamaño cuando  $n > 2$  son complicados de implementar. Generalmente los cálculos de las probabilidades de inclusión de segundo orden se vuelven rápidamente engorrosas a medida que el tamaño de la muestra se incrementa. Puede consultarse [Brewer y Hanif 1983](#) para más información.

Sin embargo, si se relaja la condición de proporcionalidad estricta entre  $\pi_k$  y  $x_k$ , se dispone del método de [Sunter 1977](#) que resulta manejable en la práctica. Este esquema asigna una probabilidad de inclusión  $\pi_k$  estrictamente proporcional a  $x_k$  para un subconjunto importante de elementos de la población. Sin embargo, a una pequeña porción de la población, que se corresponde con aquellos elementos que presentan los valores más bajos de  $x$ , se les asigna una probabilidad  $\pi_k$  igual, por simplicidad.

El esquema de Sunter es el siguiente:

- Se ordenan los elementos de la población de forma descendente por el valor de  $x$ . Sea  $k = 1, \dots, N$  el índice que identifica los elementos ordenados de esta forma. De esta forma,  $k = 1$  es el elemento con el valor más grande de  $x$ .
- Para  $k = 1$  se considera una realización de una uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon_1$ . Se calcula

$$\pi_1 = n \cdot \frac{x_1}{T_N}$$

El elemento  $k$  es seleccionado si  $\varepsilon_1 < \pi_1$ .

- Para cada elemento  $k = 2, 3, \dots, N$ , se define  $U_k = \{k, k+1, \dots, N\}$  y se considera una realización de una uniforme sobre el intervalo  $(0, 1)$ :  $\varepsilon_k$ . Se calcula

$$\pi'_k = (n - n_k) \cdot \frac{x_k}{t_k}$$

donde  $t_k = \sum_{j \in U_k} x_j$  y  $n_k$  es el número de elementos seleccionados de entre los primeros  $k-1$  elementos ordenados de la forma especificada en el primer paso. El elemento  $k$  es seleccionado si  $\varepsilon_k < \pi'_k$ .

- El proceso descrito finaliza cuando  $n_k = n$  o bien  $k = k^*$ , lo que ocurra primero, donde  $k^* = \min\{k_0, N - n + 1\}$ , con  $k_0$  igual al valor  $k$  más pequeño para el que cual se tiene  $n \cdot \frac{x_k}{t_k} \geq 1$ .



- Si  $n_{k^*} < n$  significa que el proceso no ha producido una muestra de tamaño  $n$ . El resto de elementos hasta completar la muestra ( $n - n_{k^*}$ ) son elegidos de entre los  $N - k^* + 1$  elementos mediante un muestreo aleatorio simple sin reemplazamiento de acuerdo al esquema dado por [Fan, Muller y Rezucha 1962](#) y que ya se expuso en el Ejemplo 22. Esto es, para cada elemento  $k = k^*, k^* + 1, \dots$  se considera una realización de una uniforme sobre el intervalo  $(0, 1): \varepsilon_k$ . Se calcula

$$\pi_k^0 = \frac{n - n_{k^*}}{N - k^* + 1}$$

El elemento  $k$  es seleccionado en la muestra si  $\varepsilon_k < \pi_k^0$ . El proceso concluye cuando la muestra sea de tamaño  $n$ .

Se puede demostrar que las probabilidades de inclusión de primer orden son

$$\pi_k = \begin{cases} n \cdot \frac{x_k}{T_N}, & k = 1, \dots, k^* - 1 \\ n \cdot \frac{\bar{x}_{k^*}}{T_n}, & k = k^*, \dots, N \end{cases}$$

donde

$$\bar{x}_{k^*} = \frac{t_{k^*}}{N - k^* + 1}$$

Por tanto, este esquema generalmente no conduce a un muestreo estricto con probabilidades proporcionales al tamaño, a menos que los  $N - k^* + 1$  elementos más pequeños de la variable  $x$  tomen el mismo valor.

**Comentario 20.** Este esquema satisface que  $\pi_{kl} - \pi_k \cdot \pi_l < 0$  para todo  $k \neq l$  (véase [Sunter 1977](#), para una demostración). Por tanto, el estimador de la varianza dado en (5.5) es siempre no negativo. ■

### Selección sistemática con probabilidades proporcionales al tamaño

Se presenta a continuación un mecanismo de selección de la muestra muy popular por su simplicidad, que permite obtener probabilidades de inclusión proporcionales al tamaño. Se trata de una generalización del muestreo sistemático con probabilidades iguales. El algoritmo es el siguiente:

- Se obtienen los tamaños acumulados  $T_k = T_{k-1} + x_k$ , con  $T_0 = 0$ .
- Se determina el intervalo muestral  $a$ , donde  $a$  es un entero positivo. Sea  $n$  la parte entera de  $\frac{T_N}{a}$ , entonces podemos expresar  $T_N$  como:

$$T_N = n \cdot a + c$$

donde  $0 \leq c < a$ .

- Se selecciona un número aleatorio (arranque aleatorio) entre 1 y  $a$  con igual probabilidad  $\frac{1}{a}$ :  $r$ .
- La muestra seleccionada es

$$s_r = \{k : T_{k-1} < r + (j-1) \cdot a \leq T_k, \text{ para algún } j = 1, 2, \dots, n_s\}$$

donde  $n_s$  es el tamaño de muestra, que será  $n$  si  $c < r \leq a$  o bien  $n + 1$  si  $r \leq c$ .

La selección sistemática con probabilidades proporcionales al tamaño es efectivamente un diseño *ppswor* de tamaño de muestra esencialmente fijo, ya que el tamaño será  $n$  o  $n + 1$  y las probabilidades de inclusión son

$$\pi_k = n \cdot \frac{x_k}{T_N - c}$$

suponiendo que  $n \cdot x_k \leq T_N - c = n \cdot a$  para todo  $k$ .

**Comentario 21.** La selección sistemática con probabilidades proporcionales al tamaño plantea los mismos problemas que aquellos que se estudiaron en el Tema 4 sobre el diseño *sys*: el control del tamaño de la muestra, la cuestión de la importancia de una buena ordenación de la población y el problema de la estimación de varianzas del estimador HT. ■

*Ejemplo 35.* Continuando con el Ejemplo 33, supongamos ahora que se desea seleccionar una muestra de tamaño 4 mediante el método de selección sistemática con probabilidades proporcionales al tamaño.

Dado que  $n = 4$ , el intervalo muestral  $a$  es 25 y  $c = 0$ :

$$100 = T_N = n \cdot a = 4 \cdot 25$$

Si el arranque aleatorio seleccionado entre 1 y 25 es 4, entonces los elementos correspondientes a 4, 29, 54 y 79 son aquellos que se seleccionan para la muestra. Estos elementos son 1, 3, 3 y 5, respectivamente.

Los valores observados de la variable de estudio para los tres elementos distintos seleccionados son  $y_1 = 3$ ,  $y_3 = 6$  e  $y_5 = 3$ . Por tanto, una estimación del total de  $y$  es:

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{n \cdot \frac{x_k}{T_N}} = \frac{1}{4} \cdot \left( \frac{3}{15/100} + 2 \cdot \frac{6}{30/100} + \frac{3}{15/100} \right) = 20$$

■

### 5.3.2 Muestreo con reemplazamiento

De la misma forma que se hizo en el caso de muestreo con probabilidades proporcionales al tamaño sin reemplazamiento, mostraremos también el incentivo de usar diseños tales que  $p_k \propto x_k$  cuando la selección se realiza con reemplazamiento.

#### Proposición 19

Bajo un diseño muestral con tamaño de muestra fijo  $n$ , la varianza del estimador de Hansen y Hurwitz cuando  $y_k$  es proporcional a  $p_k$ , esto es,  $y_k = c \cdot p_k$  para todo  $c \in \mathbb{R}$ , es cero.

**Demostración 19**

En efecto, podemos ver que el estimador HH es constante:

$$\hat{Y}_U^{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}} = \frac{1}{n} \sum_{i=1}^n \frac{c \cdot p_{k_i}}{p_{k_i}} = c$$

donde  $p_{k_i}$  es la probabilidad de seleccionar el elemento  $k_i$ . Por tanto, la varianza del estimador es cero.

Nuevamente, un diseño tal que  $p_k \propto y_k$  no puede ser obtenido en la práctica, ya que generalmente los valores  $y_k$  son desconocidos. Una aproximación a esta situación es encontrar un esquema de selección donde la probabilidad de extracción de cada elemento  $p_k$  sea proporcional a una variable auxiliar  $x_k$  que sea aproximadamente proporcional a  $y_k$  y cuyos valores sean positivos y conocidos para toda la población.

La condición  $p_k \propto x_k$  es equivalente a decir:

$$p_k = \frac{x_k}{\sum_{k \in U} x_k}$$

ya que  $\sum_{k \in U} p_k = 1$ .

Por definición, el estimador de Hansen y Hurwitz, bajo el diseño *ppswr*, es decir, tomando  $p_k = \frac{x_k}{\sum_{k \in U} x_k}$  es

$$\hat{Y}_U^{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}} = \left( \sum_{k \in U} x_k \right) \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}}$$

La varianza del estimador HH es:

$$\begin{aligned} \mathbb{V}(\hat{Y}_U^{HH}) &= \frac{1}{n} \cdot \sum_{k \in U} \left( \frac{y_k}{p_k} - Y_U \right)^2 \cdot p_k = \frac{1}{n} \cdot \left[ \sum_{k \in U} \frac{y_k^2}{p_k} - 2 \cdot Y_U \cdot \sum_{k \in U} y_k + Y_U^2 \cdot \sum_{k \in U} p_k \right] = \\ &= \frac{1}{n} \cdot \left[ \sum_{k \in U} \frac{y_k^2}{p_k} - 2 \cdot Y_U^2 + Y_U^2 \right] = \frac{1}{n} \cdot \left[ \sum_{k \in U} \frac{y_k^2}{p_k} - Y_U^2 \right] = \\ &= \frac{1}{n} \cdot \left[ \left( \sum_{k \in U} x_k \right) \cdot \sum_{k \in U} \frac{y_k^2}{x_k} - Y_U^2 \right] \end{aligned}$$

Se sabe que un estimador insesgado de la varianza es:

$$\begin{aligned} \hat{\mathbb{V}}(\hat{Y}_U^{HH}) &= \frac{1}{n \cdot (n-1)} \sum_{i=1}^n \left( \frac{y_{k_i}}{p_{k_i}} - \hat{Y}_U^{HH} \right)^2 = \\ &= \frac{1}{n \cdot (n-1)} \cdot \left[ \sum_{i=1}^n \left( \frac{y_{k_i}}{p_{k_i}} \right)^2 - 2 \cdot \hat{Y}_U^{HH} \cdot \sum_{i=1}^n \frac{y_{k_i}}{p_{k_i}} + n \cdot \left( \hat{Y}_U^{HH} \right)^2 \right] = \end{aligned}$$

$$\begin{aligned}
&= \frac{\left(\sum_{k \in U} x_k\right)^2}{n \cdot (n-1)} \cdot \left[ \sum_{i=1}^n \left(\frac{y_{k_i}}{x_{k_i}}\right)^2 - \frac{2}{n} \cdot \sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}} \cdot \sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}} + \frac{n}{n^2} \cdot \left(\sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}}\right)^2 \right] = \\
&= \frac{\left(\sum_{k \in U} x_k\right)^2}{n \cdot (n-1)} \cdot \left[ \sum_{i=1}^n \left(\frac{y_{k_i}}{x_{k_i}}\right)^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}}\right)^2 \right]
\end{aligned}$$

**Comentario 22.** Se puede observar que los cálculos necesarios para obtener una estimación de la varianza en el caso del estimador HH son mucho más sencillos que al usar el estimador HT. Sin embargo, a pesar de la simplicidad de los cálculos a favor del estimador HH y el diseño *ppswr*, hay que destacar que el estimador HH es normalmente menos eficiente que el de HT.

■

En cuanto a los procedimientos de selección de la muestra, un esquema sencillo para seleccionar una muestra de tamaño  $n = 1$  es usar el *método del total acumulado*, ya que los diseños *ppswor* y *ppswr* son en este caso equivalentes. Cuando el tamaño de muestra es  $n > 1$ , un mecanismo de selección de la muestra bajo muestreo con probabilidades proporcionales al tamaño con reemplazamiento consiste en repetir, de forma independiente,  $n$  veces el *método del total acumulado*.

*Ejemplo 36.* En el contexto del Ejemplo 34, considérese la población completa de 284 municipios. Una muestra de tamaño diez es extraída bajo el diseño *ppswr* con probabilidades proporcionales al valor de la población actual (variable  $x$ ) para estimar en este caso el total del valor inmobiliario de la población de acuerdo a la evaluación realizada en el año 1984 (medida en millones de coronas). Las probabilidades de inclusión se han tomado proporcionales al tamaño de la población (medida en miles). La muestra proporciona los siguientes resultados:

Población ( $x$ )	Valoración ( $y$ )
54	5246
671	59877
28	2208
62	6850
42	3773
48	4055
33	4014
446	38945
12	1162
46	4852

Sabiendo que el tamaño de la población total es de 8182 individuos, una estimación del valor total inmobiliario es:

$$\hat{Y}_U^{HH} = \left( \sum_{k \in U} x_k \right) \frac{1}{n} \sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}} = 8182 \cdot \frac{1}{10} \cdot \left( \frac{5246}{54} + \dots + \frac{4852}{46} \right) \approx 786540,34$$

Una estimación de la varianza es:

$$\begin{aligned} \hat{V}(\hat{Y}_U^{HH}) &= \frac{\left( \sum_{k \in U} x_k \right)^2}{n \cdot (n-1)} \cdot \left[ \sum_{i=1}^n \left( \frac{y_{k_i}}{x_{k_i}} \right)^2 - \frac{1}{n} \cdot \left( \sum_{i=1}^n \frac{y_{k_i}}{x_{k_i}} \right)^2 \right] = \\ &= \frac{8182^2}{10 \cdot 9} \cdot \left[ \left( \frac{5246}{54} \right)^2 + \dots + \left( \frac{4852}{46} \right)^2 - \frac{1}{10} \cdot \left( \frac{5246}{54} + \dots + \frac{4852}{46} \right)^2 \right] \approx \\ &\approx 1,149 \cdot 10^9 \end{aligned}$$

Por tanto, el  $cve(\hat{Y}_U^{HH})$  es de aproximadamente 0,043.



## Bibliografía

- Brewer, K.R.W. (1963). "A model of systematic sampling with unequal probabilities". En: *Australian Journal of Statistics* 5, págs. 5-13.
- (1975). "A simple procedure for sampling  $\pi$ pswor". En: *Australian Journal of Statistics* 17 (3), págs. 166-172.
- Brewer, K.R.W. y M. Hanif (1983). *Sampling with unequal probabilities*. Springer.
- Fan, C.T., M.E. Muller e I. Rezucha (1962). "Development of sampling plans by using sequential (item by item) techniques and digital computers". En: *Journal of the American Statistical Association* 57, págs. 387-402.
- Rao, J. N. K. (1965). "On two simple schemes of unequal probability sampling without replacement". En: *Journal of the Indian Statistical Association* 3, págs. 173-180.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
- Sen, A.R. (1953). "On the estimate of the variance in sampling with varying probabilities". En: *Journal of the Indian Society of Agricultural Statistics* 5, págs. 119-127.
- Sunter, A. B. (1977). "List sequential sampling with equal or unequal probabilities without replacement". En: *Applied Statistics* 26, págs. 261-268.
- Yates, F. y P.M. Grundy (1953). "Selection without replacement from within strata with probability proportional to size". En: *Journal of the Royal Statistical Society B* 15, págs. 253-261.

## Tema 6

### **Estimación insesgada en diseños muestrales sobre unidades elementales IV. Muestreo estratificado: definición, estimadores, varianza y estimador de la varianza. Afijación muestral óptima. Otras afijaciones bajo muestreo aleatorio simple.**

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

W.G. Cochran (1977). *Sampling Techniques*. 3rd. New York: Wiley

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

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.

## **6.1 Muestreo estratificado: definición, estimadores, varianza y estimador de la varianza**

### **6.1.1 Introducción y definición**

El muestreo estratificado es un diseño de muestreo probabilístico en el que se divide la población objeto de estudio en diferentes subpoblaciones no superpuestas, denominadas estratos. La muestra estratificada se obtiene tras seleccionar una muestra en cada estrato. La selección de la muestra en cada estrato se realiza de forma independiente.

El muestreo estratificado es un método poderoso y flexible que es ampliamente usado en la práctica. En las encuestas económicas (dirigidas a las empresas o a establecimientos), tanto coyunturales como estructurales, es el tipo de diseño que se suele utilizar. Por ejemplo, en *Índices de Comercio al por Menor*<sup>1</sup> y la *Estadística Estructural de Empresas: Sectores Industrial, Comercio, Servicios*<sup>2</sup>.

<sup>1</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176900&menu=metodologia&idp=1254735576799](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176900&menu=metodologia&idp=1254735576799)

<sup>2</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736143952&menu=metodologia&idp=1254735576550](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736143952&menu=metodologia&idp=1254735576550)

Algunas variables que se usan para realizar la partición de la población son, entre otras, la actividad económica, principal o no principal, en función de agrupaciones de CNAE<sup>3</sup>, el tamaño en función del número de asalariados y la Comunidad Autónoma donde se ubica la sede de la empresa o del establecimiento, información que se encuentra disponible en el marco usado para las encuestas económicas, el *Directorio Central de Empresas (DIRCE)*<sup>4</sup>. El muestreo estratificado también es utilizado en encuestas a hogares. Por ejemplo, en la *Encuesta de Población Activa*<sup>5</sup> se aplica un diseño bietápico, con estratificación de las unidades de primera etapa (las secciones censales) y sin submuestreo en las unidades secundarias (viviendas familiares principales). Las secciones censales, división territorial con fines estadísticos y administrativos, son áreas geográficas perfectamente delimitadas cuyo tamaño de población viene regulado por la Ley Orgánica del Régimen Electoral General<sup>6</sup>.

A continuación se detallan algunas de las razones por las que el muestreo estratificado es tan popular:

1. Cuando se requieren estimaciones de precisión separadas para determinadas subpoblaciones (dominios de estudio) y la pertenencia al dominio de cada elemento de la población aparece definido en el marco muestral, entonces cada dominio de estudio puede ser tratado como un estrato separado y obtener así una muestra probabilística adecuada de cada estrato. Por ejemplo, algunos estratos pueden ser exhaustivos, todas sus unidades pertenecen a la muestra, mientras que otros serán muestrales.
2. La tasa de falta de respuesta, los errores de medida y la información auxiliar pueden diferir considerablemente de una subpoblación a otra. Si ocurre esto, parece adecuado pensar que la elección de un diseño de muestreo y un estimador quizá no debería ser el mismo en todos los estratos y debería elegirse el más conveniente en cada subpoblación para así mejorar la eficiencia en la estimación.
3. La conveniencia administrativa puede imponer el uso de la estratificación. Por ejemplo, si la agencia encargada de la encuesta dispone de oficinas de campo en una serie de distritos geográficos, cada una de las cuales puede supervisar la encuesta para una parte de la población. En este caso, parece natural tomar cada distrito como un estrato.
4. Cuando la estratificación permite dividir la población heterogénea en subpoblaciones internamente homogéneas. Si cada estrato es homogéneo internamente, en el sentido de que las medidas de las características de estudio varían poco de una unidad a otra, se puede obtener una estimación precisa del parámetro poblacional de estudio para cualquier estrato a partir de una pequeña muestra en

---

<sup>3</sup>[https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614](https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614)

<sup>4</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550)

<sup>5</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176918&menu=metodologia&idp=1254735976595](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=metodologia&idp=1254735976595)

<sup>6</sup><https://www.boe.es/eli/es/lo/1985/06/19/5>

ese estrato. Esto producirá una ganancia en precisión en las estimaciones finales de los parámetros poblacionales de interés.

El muestreo estratificado tiene la ventaja de producir muestras más representativas<sup>7</sup> de la población respecto de la variable o variables de estratificación, debido a que se asegura que todos los elementos de cada subpoblación tienen representación en la muestra. Por tanto, la muestra se extiende sobre todos los estratos definidos. Si la estratificación se realiza de forma correcta, tal que suponga la formación de grupos internamente homogéneos y heterogéneos entre ellos en relación a las características de estudio, entonces se produce una ganancia en precisión en la estimación de dichas características para la población completa.

Por otra parte, el muestreo estratificado permite aplicar distintos diseños de muestreo en cada subpoblación, así como un estimador diferente, por lo que presenta la ventaja de poder usar el más apropiado para cada estrato. Por último, si el muestreo estratificado se organiza y administra convenientemente, este tipo de diseño puede suponer la reducción del coste global de la encuesta.

El principal inconveniente del muestreo estratificado es que en el marco debe estar disponible información auxiliar para realizar la estratificación de la población para todas las unidades. Además, esta información auxiliar también debe de estar depurada y reflejar lo más fielmente la realidad, ya que, en caso contrario, la estimación se verá afectada y será necesario realizar un proceso de postestratificación<sup>8</sup> para poder obtener estimaciones acuradas.

### Consideraciones técnicas para aplicar un muestreo estratificado

Si el estadístico encargado de la investigación decide realizar un muestreo estratificado, debe tener en cuenta las siguientes cuestiones:

- i. Construcción de los estratos
  - (a) El estadístico debe elegir, en caso de que sea posible, la característica o características que va a usar para dividir la población en estratos. Las características escogidas se denominan variables de estratificación. Por ejemplo, podría considerarse la edad y el sexo como variables de estratificación o bien estratificar por grupos ocupacionales.
  - (b) Decidir cuántos estratos debería haber. Por ejemplo, si se usa como variable de estratificación la edad, se debe determinar el número de grupos de edad.
  - (c) Determinar los límites de cada estrato a partir de las variables de estratificación elegidas. Por ejemplo, si se usan grupos de edad, se debe decidir qué

---

<sup>7</sup>se entiende representativo como la propiedad deseable de una estrategia muestral (diseño + estimador) que conduce a estimaciones acuradas

<sup>8</sup>véase la sección 7.6 del ([Särndal, Swensson y Wretman 1992](#))



intervalos de edad se usarán para configurar los estratos.

En ocasiones la construcción de los estratos puede venir fijada por los dominios de difusión, que a su vez serán especificados por las necesidades de información internas (de otros ministerios o subdirecciones dentro del INE) y externas (Reglamentos/Directivas europeos, etc.).

- ii. Elección del diseño de muestreo y el estimador dentro de cada estrato
  - (a) Especificación del diseño de muestreo y el tamaño de muestra en cada estrato. A menudo se suele aplicar el mismo tipo de diseño en todos los estratos.
  - (b) Especificación del estimador a usar en cada estrato. Se suele aplicar también el mismo en todos los estratos.

### Notación

Considérese una población constituida por  $N$  elementos  $\{u_1, \dots, u_N\}$ , que se denota por  $U = \{1, \dots, N\}$  y supóngase que el parámetro de interés es el total poblacional de la *variable de estudio*  $y$ . Se realiza una estratificación de la población  $U$ , es decir, se realiza una partición de la población en  $H$  subpoblaciones, denominados estratos, denotados por  $U_1, U_2, \dots, U_H$ . Por tanto,  $U_h$  contiene los elementos de la población que pertenecen al estrato  $h$ ,  $\forall h = 1, \dots, H$ .

El muestreo estratificado consiste en seleccionar, en cada estrato  $h$ , una muestra probabilística de  $U_h$  de acuerdo al diseño de muestreo establecido para ese estrato,  $p_h(\cdot)$ ,  $h = 1, \dots, H$ . El resultado de la muestra total será la unión de todas las muestras seleccionadas en cada estrato, esto es,

$$s = s_1 \cup s_2 \cup \dots \cup s_H$$

El número de unidades seleccionadas de  $U_h$  se denomina el tamaño muestral del estrato  $h$  y se denota por  $n_h$ . Así, el tamaño muestral total,  $n$ , puede representarse como

$$n = \sum_{h=1}^H n_h$$

Como consecuencia de la independencia en la selección de la muestra en cada estrato, el diseño del muestreo estratificado es

$$p(s) = p_1(s_1)p_2(s_2) \cdots p_H(s_H)$$

Se denota diseño  $st$ .

El número de elementos en el estrato  $h$  se supone conocido y se denota por  $N_h$ . Dado que los estratos son una partición de la población total, se tiene:

$$N = \sum_{h=1}^H N_h$$

Por otra parte, el total poblacional de la variable  $y$  puede representarse como sigue:

$$Y_U = \sum_{k \in U} y_k = \sum_{h=1}^H Y_{U_h} = \sum_{h=1}^H N_h \cdot \bar{y}_{U_h}$$

donde  $Y_{U_h} = \sum_{k \in U_h} y_k$  es el total del estrato  $h$  y  $\bar{y}_{U_h} = \frac{1}{N_h} \sum_{k \in U_h} y_k$  la media del estrato  $h$ .

La media poblacional de la variable  $y$  puede representarse como

$$\bar{y}_U = \sum_{h=1}^H W_h \cdot \bar{y}_{U_h}$$

donde  $W_h = \frac{N_h}{N}$  es el peso poblacional del estrato  $h$ .

El número de elementos de la población  $U$  que pertenecen a un determinado dominio  $d$ , se puede expresar como

$$N_d = \sum_{h=1}^H \sum_{k \in U_h} z_{dk}$$

donde

$$z_{dk} = \begin{cases} 1, & \text{si } k \in U_d \\ 0, & \text{si } k \notin U_d \end{cases} \quad \text{para todo } k \in 1, \dots, N.$$

Esto es, una variable que toma únicamente dos valores, 0 y 1, en función de si el elemento pertenece al dominio o no. Sea  $N_{dh} = \sum_{k \in U_h} z_{dk}$  el número de elementos del estrato  $U_h$  que pertenecen al dominio  $d$  y  $P_{dh} = \frac{1}{N_h} \sum_{k \in U_h} z_{dk} = \frac{N_{dh}}{N_h}$  la proporción de elementos de dicho estrato que pertenecen al dominio, entonces  $N_d$  se puede escribir como

$$N_d = \sum_{h=1}^H N_{dh} = \sum_{h=1}^H N_h \cdot P_{dh}$$

Por último, la proporción poblacional de los elementos de  $U$  que pertenecen a un determinado dominio  $d$  se representa como

$$P_d = \frac{N_d}{N} = \sum_{h=1}^H W_h \cdot P_{dh}$$

### 6.1.2 Estimadores, varianza y estimador de la varianza

Supongamos que se considera usar el estimador de Horvitz-Thompson en todos los estratos, mientras que los diseños de muestreo aplicados sí podrían ser diferentes de una subpoblación a otra. A continuación se obtendrá el estimador de Horvitz-Thompson bajo el diseño del muestreo estratificado, así como la varianza del estimador y un estimador de dicha varianza.

**Proposición 20**

Bajo muestreo estratificado, el estimador de Horvitz-Thompson (HT) del total poblacional  $Y_U = \sum_{k \in U} y_k$  es

$$\hat{Y}_U^{\text{HT}} = \sum_{h=1}^H \hat{Y}_{U_h}^{\text{HT}} \quad (6.1)$$

donde  $\hat{Y}_{U_h}^{\text{HT}}$  es el estimador HT del total poblacional del estrato  $h$ ,  $Y_{U_h}$ .

La varianza del estimador es

$$\mathbb{V}_{st}(\hat{Y}_U^{\text{HT}}) = \sum_{h=1}^H \mathbb{V}_{p_h}(\hat{Y}_{U_h}^{\text{HT}}) \quad (6.2)$$

donde  $\mathbb{V}_{p_h}(\hat{Y}_{U_h}^{\text{HT}})$  es la varianza del estimador  $\hat{Y}_{U_h}^{\text{HT}}$  bajo el diseño de muestreo  $p_h$  utilizado en el estrato  $h$ .

Un estimador insesgado de la varianza viene dado por

$$\hat{\mathbb{V}}_{st}(\hat{Y}_U^{\text{HT}}) = \sum_{h=1}^H \hat{\mathbb{V}}_{p_h}(\hat{Y}_{U_h}^{\text{HT}}) \quad (6.3)$$

suponiendo que existe un estimador insesgado de  $\mathbb{V}_{p_h}(\hat{Y}_{U_h}^{\text{HT}})$ , que es  $\hat{\mathbb{V}}_{p_h}(\hat{Y}_{U_h}^{\text{HT}})$ .

**Demostración 20**

Por definición de muestreo estratificado, es fácil ver que para todo  $k \in U_h$ , se tiene

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(k \in s_h)$$

Por tanto, el estimador HT del total poblacional es

$$\hat{Y}_U^{\text{HT}} = \sum_{h=1}^H \sum_{k \in s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^H \hat{Y}_{U_h}^{\text{HT}}$$

Dado que la muestra estratificada se obtiene, por definición, seleccionando una muestra en cada estrato de forma independiente, las variables aleatorias  $\hat{Y}_{U_h}^{\text{HT}}$  son independientes y, de esta forma, se obtienen de forma sencilla los resultados (6.2) y (6.3).

**Comentario 23.** El muestreo estratificado permite aplicar diferentes diseños de muestreo y estimadores en los distintos estratos. Sin embargo, es habitual usar el mismo diseño y el mismo estimador en todas las subpoblaciones. Uno de los diseños más populares es

aplicar la selección de una muestra aleatoria simple en cada estrato. Denotaremos a este diseño por *strs* y se denomina *muestreo aleatorio estratificado*.

### Proposición 21

Bajo el diseño del muestreo aleatorio estratificado, el estimador HT del total poblacional es

$$\hat{Y}_U^{\text{HT}} = \sum_{h=1}^H N_h \cdot \bar{y}_{s_h}$$

donde  $\bar{y}_{s_h} = \sum_{k \in s_h} \frac{y_k}{n_h}$  es la media muestral del estrato  $h$ .

La varianza del estimador HT es

$$\mathbb{V}_{\text{strs}} \left( \hat{Y}_U^{\text{HT}} \right) = \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{U_h}}^2$$

donde  $f_h = \frac{n_h}{N_h}$  es la fracción de muestreo en el estrato  $h$  y

$$S_{y_{U_h}}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2$$

es la cuasivarianza del estrato  $h$  y  $\bar{y}_{U_h} = \sum_{k \in U_h} \frac{y_k}{N_h}$  es la media de la población  $U_h$ .

Un estimador insesgado de la varianza viene dado por

$$\hat{\mathbb{V}}_{\text{strs}} \left( \hat{Y}_U^{\text{HT}} \right) = \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{s_h}}^2$$

donde

$$S_{y_{s_h}}^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_{s_h})^2$$

es la cuasivarianza muestral del estrato  $h$ .

### Demostración 21

Recordemos los resultados obtenidos en el Tema 3 para el diseño aleatorio simple sin reemplazamiento:

- Estimador HT:

$$\hat{Y}_{U_h}^{\text{HT}} = N_h \sum_{k \in s_h} \frac{y_k}{n_h} = N_h \cdot \bar{y}_{s_h}$$

- Varianza:

$$\mathbb{V}_{srswor} \left( \hat{Y}_{U_h}^{HT} \right) = N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{U_h}}^2$$

- Estimador de la varianza:

$$\hat{\mathbb{V}}_{srswor} \left( \hat{Y}_{U_h}^{HT} \right) = N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{s_h}}^2$$

Utilizando estos resultados en las expresiones (6.1), (6.2) y (6.3), se llega a lo que se pretendía demostrar.

*Ejemplo 37.* Se dispone de una población constituida por 284 municipios (Apéndice B de Särndal, Swensson y Wretman 1992) de la que se desea estimar el total de escaños del Partido Socialdemócrata en el consejo municipal (variable  $y$ ). La población se ha dividido en cuatro estratos en función de la variable de estratificación número de escaños. Se realiza un muestreo estratificado donde en cada estrato se selecciona una muestra aleatoria simple independiente de municipios. Los resultados obtenidos son los siguientes:

Número de escaños	$N_h$	$n_h$	$\sum_{k \in s_h} y_k$	$\sum_{k \in s_h} y_k^2$
31 – 40	44	5	89	1647
41 – 50	168	21	441	9735
51 – 70	56	10	280	8294
71 o más	16	4	152	5794

Bajo el diseño  $strs$ , el estimador de HT del total de escaños del Partido Socialdemócrata es

$$\hat{Y}_U^{HT} = \sum_{h=1}^H N_h \cdot \bar{y}_{s_h} = 44 \cdot \frac{89}{5} + 168 \cdot \frac{441}{21} + 56 \cdot \frac{280}{10} + 16 \cdot \frac{152}{4} = 6487,20$$

Calculemos ahora las cuasivarianzas muestrales de cada estrato:

$$S_{y_{s_h}}^2 = \frac{1}{n_h - 1} \left( \sum_{k \in s_h} y_k^2 - n_h \cdot \bar{y}_{s_h}^2 \right)$$

$$S_{y_{s_1}}^2 = 15,7; \quad S_{y_{s_2}}^2 = 23,7; \quad S_{y_{s_3}}^2 = \frac{454}{9}; \quad S_{y_{s_4}}^2 = 6$$

La estimación de la varianza es

$$\begin{aligned} \hat{\mathbb{V}}_{strs} \left( \hat{Y}_U^{HT} \right) &= \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{s_h}}^2 = 44^2 \cdot \frac{1 - 5/44}{5} \cdot 15,7 + \\ &+ 168^2 \cdot \frac{1 - 21/168}{21} \cdot 23,7 + 56^2 \cdot \frac{1 - 10/56}{10} \cdot \frac{454}{9} + \end{aligned}$$

$$+ 16^2 \cdot \frac{1 - 4/16}{4} \cdot 6 \approx 46541,93$$

La estimación del error relativo o coeficiente de variación del estimador es aproximadamente de 3.33 %.

$$cve(\hat{Y}_U^{\text{HT}}) = \frac{\hat{\sigma}_{\text{strs}}(\hat{Y}_U^{\text{HT}})}{\hat{Y}_U^{\text{HT}}} \approx 0,0333$$

■

### Estimación de la media poblacional bajo el diseño str

Un estimador insesgado de la media poblacional  $\bar{y}_U$  se obtiene dividiendo el estimador de Horvitz-Thompson para el caso del total por el tamaño poblacional  $N$ :

$$\hat{\bar{y}}_U^{\text{HT}} = \frac{1}{N} \cdot \hat{Y}_U^{\text{HT}} = \frac{1}{N} \sum_{h=1}^H N_h \cdot \bar{y}_{s_h} = \sum_{h=1}^H W_h \cdot \bar{y}_{s_h}$$

La varianza y el estimador de la varianza pueden obtenerse de forma sencilla:

$$\begin{aligned} \mathbb{V}_{\text{strs}}(\hat{\bar{y}}_U^{\text{HT}}) &= \frac{1}{N^2} \cdot \mathbb{V}[\hat{Y}_U^{\text{HT}}] = \sum_{h=1}^H W_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{U_h}}^2 \\ \hat{\mathbb{V}}_{\text{strs}}(\hat{\bar{y}}_U^{\text{HT}}) &= \sum_{h=1}^H W_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{y_{s_h}}^2 \end{aligned}$$

### Estimación de la proporción y el número de elementos que pertenecen a un dominio bajo el diseño str

Sea  $n_{dh} = \sum_{k \in s_h} z_{dk}$  el número de elementos de la muestra extraída en el estrato  $h$  que pertenecen al dominio y  $p_{dh} = \frac{n_{dh}}{n_h}$  la proporción de elementos de la muestra extraída en el estrato  $h$  que pertenecen al dominio. Dado que la variable  $z_d$  es un caso particular de una característica que toma únicamente los valores 0 y 1, entonces el estimador HT es, aplicando lo estudiado,

$$\hat{N}_d = \sum_{h=1}^H N_h \cdot p_{dh}$$

Tomando  $Q_{dh} = 1 - P_{dh}$ , la varianza del estimador viene dada por:

$$\begin{aligned} \mathbb{V}_{\text{strs}}(\hat{N}_d) &= \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{z_d U_h}^2 = \sum_{h=1}^H N_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{\sigma_{z_d U_h}^2}{n_h} = \\ &= \sum_{h=1}^H N_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_{dh} \cdot Q_{dh}}{n_h} \end{aligned}$$

ya que

$$\sigma_{z_d U_h}^2 = P_{dh} \cdot (1 - P_{dh}) = P_{dh} \cdot Q_{dh}$$

Un estimador insesgado de la varianza es:

$$\hat{\mathbb{V}}_{strs}(\hat{N}_d) = \sum_{h=1}^H N_h^2 \cdot (1 - f_h) \cdot \frac{S_{z_d s_h}^2}{n_h} = \sum_{h=1}^H N_h^2 \cdot (1 - f_h) \cdot \frac{p_{dh} \cdot q_{dh}}{n_h - 1}$$

ya que  $S_{z_d s_h}^2 = \frac{n_h}{n_h - 1} \cdot p_{dh} \cdot q_{dh}$ , donde  $q_{dh} = 1 - p_{dh}$

De forma análoga, para el caso de la proporción de individuos que pertenecen al dominio, un estimador insesgado para  $P_d$  es

$$\hat{P}_d = \sum_{h=1}^H W_h \cdot p_{dh}$$

La varianza y el estimador de la varianza se pueden calcular a partir de las expresiones vistas para el estimador de  $N_d$ :

$$\begin{aligned} \mathbb{V}_{strs}(\hat{P}_d) &= \frac{1}{N^2} \cdot \mathbb{V}_{strs}(\hat{N}_d) = \sum_{h=1}^H W_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_{dh} \cdot Q_{dh}}{n_h} \\ \hat{\mathbb{V}}_{strs}(\hat{P}_d) &= \frac{1}{N^2} \cdot \hat{\mathbb{V}}_{strs}(\hat{N}_d) = \sum_{h=1}^H W_h^2 \cdot (1 - f_h) \cdot \frac{p_{dh} \cdot q_{dh}}{n_h - 1} \end{aligned}$$

## 6.2 Afijación óptima

Consideremos una población dividida en  $H$  estratos de la cual se desea estimar el total poblacional de la variable  $y$ . Se decide realizar un muestreo estratificado para el cual ya se han decidido los diseños a aplicar en cada estrato. El estimador a utilizar es el de Horvitz-Thompson. Supongamos que el diseño final produce una varianza del estimador de la forma:

$$\mathbb{V}_{st}(\hat{Y}_U^{HT}) = \sum_{h=1}^H \frac{A_h}{n_h} + B = V \quad (6.4)$$

donde  $A_h$  y  $B$  no dependen del tamaño muestral  $n_h$ ,  $h = 1, \dots, H$ . Antes de seleccionar la muestra en cada estrato, el estadístico debe determinar cuántas unidades muestrales  $n_h$  extraerá de cada estrato. El problema de determinar  $n_h$  se conoce como el problema de la *afijación de la muestra*.

### Definición 16

El concepto de *afijación de la muestra* se define como el reparto o distribución del tamaño de la muestra total,  $n$ , entre los distintos estratos.

**Comentario 24.** La varianza del estimador HT bajo el diseño *strs* es de la forma dada en (6.4). En efecto:

$$\mathbb{V}_{strs}(\hat{Y}_U^{HT}) = \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{yU_h}^2 = \sum_{h=1}^H \frac{N_h^2 \cdot S_{yU_h}^2}{n_h} - \sum_{h=1}^H N_h \cdot S_{yU_h}^2$$

donde se toma  $A_h = N_h^2 \cdot S_{yU_h}^2$  y  $B = - \sum_{h=1}^H N_h \cdot S_{yU_h}^2$ .

■

Además, supongamos que el coste total de la encuesta puede ser expresado como

$$C = c_0 + \sum_{h=1}^H n_h \cdot c_h$$

donde  $c_0$  es el coste fijo y  $c_h$  es el coste de observar un elemento del estrato  $h$ .

### Definición 17

El problema de la afijación óptima de la muestra se puede formular de dos formas distintas:

1. Determinación de los tamaños muestrales  $n_h$  que minimicen la varianza del estimador  $V$  sujeto a la restricción de un coste fijo  $C$ .

$$\begin{cases} \text{minimizar } V = \sum_{h=1}^H \frac{A_h}{n_h} + B \\ \text{sujeto a } C = c_0 + \sum_{h=1}^H n_h \cdot c_h \end{cases} \quad (6.5)$$

2. Determinación de los tamaños muestrales  $n_h$  que minimicen el coste total de la encuesta  $C$  sujeto a una precisión fijada  $V$ .

$$\begin{cases} \text{minimizar } C = c_0 + \sum_{h=1}^H n_h \cdot c_h \\ \text{sujeto a } V = \sum_{h=1}^H \frac{A_h}{n_h} + B \end{cases} \quad (6.6)$$

### Proposición 22

En el diseño de muestreo estratificado para el cual la varianza del estimador HT es de la forma (6.4), la *afijación óptima de la muestra*, suponiendo la función de costes

$C = c_0 + \sum_{h=1}^H n_h \cdot c_h$ , es tal que  $n_h$  es proporcional a  $\left(\frac{A_h}{c_h}\right)^{1/2}$ , esto es:

$$n_h \propto \left(\frac{A_h}{c_h}\right)^{1/2}$$

### Demostración 22



El problema de optimización que se plantea es el especificado en (6.5) o en (6.6). Se resuelve de forma análoga en los dos casos.

Sea  $V^* = V - B$  y  $C^* = C - c_0$ , entonces el problema es equivalente a minimizar el producto

$$V^* \cdot C^* = \left( \sum_{h=1}^H \frac{A_h}{n_h} \right) \cdot \left( \sum_{h=1}^H n_h \cdot c_h \right)$$

De la desigualdad de Cauchy se tiene

$$\left( \sum a_h^2 \right) \cdot \left( \sum b_h^2 \right) \geq \left( \sum a_h \cdot b_h \right)^2.$$

La igualdad se da si y solo si  $\frac{b_h}{a_h}$  es constante para cada  $h$ . Tomando  $a_h = \left( \frac{A_h}{n_h} \right)^{1/2}$  y  $b_h = (n_h \cdot c_h)^{1/2}$  se llega a

$$V^* \cdot C^* \geq \left[ \sum_{h=1}^H (A_h \cdot c_h)^{1/2} \right]^2$$

La igualdad se da cuando  $\left( \frac{n_h \cdot c_h}{\frac{A_h}{n_h}} \right)^{1/2} \equiv \text{constante}$  o, de forma equivalente, se puede decir que

$$n_h \propto \left( \frac{A_h}{c_h} \right)^{1/2}$$

Si se plantea el problema de **minimizar la varianza del estimador HT para un coste C fijado**, esto es, el problema especificado en (6.5), entonces dado que  $n_h = \alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2}$  para una constante  $\alpha$  y  $C = c_0 + \sum_{h=1}^H n_h \cdot c_h$ , se tiene:

$$C = C_0 + \sum_{h=1}^H n_h \cdot c_h = c_0 + \sum_{h=1}^H \alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2} \cdot c_h = c_0 + \alpha \sum_{h=1}^H (A_h \cdot c_h)^{1/2}$$

Por tanto

$$\alpha = \frac{C - c_0}{\sum_{h=1}^H (A_h \cdot c_h)^{1/2}}$$

El tamaño del estrato  $h$  viene dado por la expresión

$$n_h = \frac{C - c_0}{\sum_{h=1}^H (A_h \cdot c_h)^{1/2}} \cdot \left( \frac{A_h}{c_h} \right)^{1/2}, \quad h = 1, \dots, H.$$

Usando lo anterior, la varianza mínima es

$$V_{opt} = \frac{1}{C - c_0} \cdot \left[ \sum_{h=1}^H (A_h \cdot c_h)^{1/2} \right]^2 + B$$

Si se plantea el problema de **minimizar el coste C fijado para una precisión dada V**, esto es, el problema especificado en (6.6), entonces dado que  $n_h = \alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2}$  y

$V = \sum_{h=1}^H \frac{A_h}{n_h} + B$ , se tiene:

$$V = \sum_{h=1}^H \frac{A_h}{\alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2}} + B = \frac{1}{\alpha} \sum_{h=1}^H (A_h \cdot c_h)^{1/2} + B$$

Por tanto

$$\alpha = \frac{\sum_{h=1}^H (A_h \cdot c_h)^{1/2}}{V - B}$$

El tamaño del estrato  $h$  viene dado por

$$n_h = \frac{\sum_{h=1}^H (A_h \cdot c_h)^{1/2}}{V - B} \cdot \left( \frac{A_h}{c_h} \right)^{1/2}, \quad h = 1, \dots, H.$$

El coste mínimo es

$$C_{opt} = c_0 + \frac{1}{V - B} \cdot \left[ \sum_{h=1}^H (A_h \cdot c_h)^{1/2} \right]^2$$

**Comentario 25.** Sea el tamaño de la muestra total un valor fijo dado,  $n$ . Dado que  $n = \sum_{h=1}^H n_h$ , es posible expresar la afijación óptima para los problemas de optimización planteados en (6.5) y (6.6) como:

$$n_h = n \cdot \frac{\left( \frac{A_h}{c_h} \right)^{1/2}}{\sum_{h=1}^H \left( \frac{A_h}{c_h} \right)^{1/2}} \quad (6.7)$$

En efecto, como  $n_h = \alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2}$ , se tiene:

$$n = \sum_{h=1}^H \alpha \cdot \left( \frac{A_h}{c_h} \right)^{1/2}; \quad \alpha = \frac{n}{\sum_{h=1}^H \left( \frac{A_h}{c_h} \right)^{1/2}}; \quad n_h = \frac{n}{\sum_{h=1}^H \left( \frac{A_h}{c_h} \right)^{1/2}} \cdot \left( \frac{A_h}{c_h} \right)^{1/2}$$

■

Bajo el diseño *strs*, como ya se comentó en el Comentario 24, la varianza del estimador HT se puede expresar de la forma

$$\mathbb{V}_{strs}(\hat{Y}_U^{HT}) = \sum_{h=1}^H \frac{A_h}{n_h} + B$$

con  $A_h = N_h^2 \cdot S_{yU_h}^2$  y  $B = - \sum_{h=1}^H N_h \cdot S_{yU_h}^2$ . Por tanto, la afijación óptima para el caso de minimizar la varianza del estimador dado un coste fijado  $C$  o bien minimizar el coste  $C$  dada una precisión  $V$ , usando (6.7) para un tamaño de muestra global  $n$ , es

$$n_h = n \cdot \frac{\frac{N_h \cdot S_{yU_h}}{c_h^{1/2}}}{\sum_{h=1}^H \frac{N_h \cdot S_{yU_h}}{c_h^{1/2}}} \quad (6.8)$$

**Comentario 26.** Se puede observar que cuanto mayor es la variación dentro de un estrato y mayor es el tamaño poblacional de dicho estrato, mayor será el número de unidades muestrales del estrato  $h$ . Asimismo, cuanto menor es el coste de observar una unidad en el estrato  $h$ , mayor será  $n_h$ . ■

**Comentario 27.** La afijación óptima podría producir un tamaño de muestra  $n_h$  para algún estrato mayor que el tamaño poblacional correspondiente de dicho estrato,  $N_h$ . Este problema puede ocurrir en la práctica y surge cuando la fracción de muestreo total es considerable y algunos estratos son mucho más variables que otros. A continuación se detalla el algoritmo a seguir cuando ocurre este problema para finalmente tener una afijación óptima tal que  $n_h \leq N$ ,  $\forall h = 1, \dots, H$ . Se ilustrará para el caso en que  $n$  es un valor fijo dado:

- Sea  $H'$  el conjunto de estratos para los que se obtiene  $n_h > N_h$ .
- Tomar

$$\begin{aligned} n'_h &= N_h, \quad \forall h \in H' \\ n'_h &= \left( n - \sum_{k \in H'} N_k \right) \cdot \frac{\left( \frac{A_h}{c_h} \right)^{1/2}}{\sum_{h \notin H'} \left( \frac{A_h}{c_h} \right)^{1/2}}, \quad \forall h \notin H' \end{aligned}$$

- Si todos los estratos verifican ahora  $n_h \leq N_h$ , ésta será la solución óptima. En caso contrario, se repite el proceso descrito hasta que se cumpla la condición  $n_h \leq N_h \quad \forall h$ . ■

### 6.2.1 Afijación en el caso de múltiples variables de estudio

Sean  $y_1, y_2, \dots, y_I$  una serie de variables de estudio con  $I \geq 2$ , de las que se quiere estimar el total poblacional de cada una de ellas,  $\sum_{k \in U} y_{ik}, \forall i = 1, \dots, I$ .

La mejor afijación para una variable no será en general la mejor para otra variable distinta. Por tanto, alguna solución debe ser adoptada en una encuesta donde existen numerosas características de estudio para obtener la mejor afijación. El primer paso será reducir el conjunto de variables consideradas para la obtención de la afijación a un número relativamente pequeño de características que se consideran más importantes.

Una posible forma de afijación es la sugerida por [Yates 1960](#), método que se describe a continuación. Considerando el diseño *strs*, el estimador del total de la variable  $y_i$  viene dado por

$$\hat{Y}_{iU}^{\text{HT}} = \sum_{h=1}^H N_h \cdot \bar{y}_{is_h}$$

donde  $\bar{y}_{is_h}$  es la media de la variable  $y_i$  en la muestra  $s_h$  extraída del estrato  $h$ . La varianza es

$$\mathbb{V}_i = \mathbb{V}_{\text{strs}} \left( \hat{Y}_{iU}^{\text{HT}} \right) = B_i + \sum_{h=1}^H \frac{A_{ih}}{n_h}$$

con

$$A_{ih} = N_h^2 \cdot S_{iU_h}^2; \quad B_i = - \sum_{h=1}^H N_h \cdot S_{iU_h}^2$$

donde  $S_{iU_h}^2$  es la cuasivarianza poblacional de la variable  $y_i$  en el estrato  $U_h$ .

Consideremos la siguiente combinación lineal de las varianzas  $\mathbb{V}_i$  dada por

$$\mathbb{V}_{\text{lin}} = \sum_{i=1}^I \omega_i \cdot \mathbb{V}_i = \sum_{i=1}^I \omega_i \cdot B_i + \sum_{h=1}^H \sum_{i=1}^I \frac{\omega_i \cdot A_{ih}}{n_h}$$

donde  $\omega_i$  es el peso que representa la importancia de la variable  $y_i$ .

El método consiste en minimizar la expresión  $\mathbb{V}_{\text{lin}}$  sujeto a la función de costes

$$C = c_0 + \sum_{h=1}^H n_h \cdot c_h,$$

o bien minimizar el coste sujeto a una varianza  $\mathbb{V}_{\text{lin}}$  fijada. Este procedimiento tiene la ventaja de simplificar el problema reduciéndolo al caso unidimensional.

Sea  $V^* = \sum_{h=1}^H \sum_{i=1}^I \frac{\omega_i \cdot A_{ih}}{n_h}$  y  $C^* = C - c_0$ , entonces el problema planteado es equivalente a minimizar el producto

$$V^* \cdot C^* = \left( \sum_{h=1}^H \sum_{i=1}^I \frac{\omega_i \cdot A_{ih}}{n_h} \right) \cdot \left( \sum_{h=1}^H n_h \cdot c_h \right)$$

De la desigualdad de Cauchy se tiene

$$\left( \sum a_h^2 \right) \cdot \left( \sum b_h^2 \right) \geq \left( \sum a_h \cdot b_h \right)^2.$$

donde  $a_h = \left( \sum_{i=1}^I \frac{\omega_i \cdot A_{ih}}{n_h} \right)^{1/2}$  y  $b_h = (n_h \cdot c_h)^{1/2}$ .

La igualdad se da cuando

$$\frac{b_h}{a_h} = \frac{(n_h \cdot c_h)^{1/2}}{\left( \sum_{i=1}^I \frac{\omega_i \cdot A_{ih}}{n_h} \right)^{1/2}} \equiv \text{constante}$$

o, de forma equivalente, se puede decir que, bajo el diseño *strs*,

$$n_h \propto \frac{1}{\sqrt{c_h}} \cdot \left( \sum_{i=1}^I \omega_i \cdot A_{ih} \right)^{1/2} = \frac{N_h}{\sqrt{c_h}} \cdot \left( \sum_{i=1}^I \omega_i \cdot S_{iU_h}^2 \right)^{1/2}$$

La principal debilidad de este procedimiento es la posible arbitrariedad en la elección de los pesos  $\omega_i$  de cada variable. Para más información, véase (Rao 1979).

Otra forma de abordar el problema de afijación consiste en minimizar el coste dado por la función  $C = c_0 + \sum_{h=1}^H n_h \cdot c_h$  bajo las siguientes restricciones:

- Se especifican unas tolerancias para la varianza correspondiente a cada variable.

$$\mathbb{V}_i \leq \mathbb{V}_{i0}, \quad i = 1, \dots, I.$$

donde  $\mathbb{V}_{i0}$  es la varianza deseada para la variable  $i$

- $n_h \leq Nh$ ,  $h = 1, \dots, H$ .
- $n_h \geq 1$ ,  $h = 1, \dots, H$  (requerida en caso de querer calcular la media de un estrato) o bien  $n_h \geq 2$ ,  $h = 1$  (requerida en caso de querer calcular la cuasivarianza en un estrato).

Este problema de optimización puede ser descrito como un problema de programación matemática convexa. En Danielsson 1975 se demuestra que este problema tiene una solución que puede ser enunciada analíticamente aunque en una forma compleja.

### 6.3 Otras afijaciones bajo muestreo aleatorio simple

En este apartado se van a estudiar otros tipos de afijaciones alternativas a la afijación óptima con costes distintos de observar una unidad en cada estrato, aplicándolas todas al caso del diseño del muestreo aleatorio estratificado (diseño *strs*). A partir de ahora supondremos iguales los costes de observar una unidad en todos los estratos,  $c_h \equiv c, \forall h = 1, \dots, H$ .

#### 6.3.1 Afijación de Neyman

La afijación óptima teniendo en cuenta que los costes de observar una unidad son iguales en todos los estratos viene dado por

$$n_h = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^H N_h \cdot S_{yU_h}} \quad (6.9)$$

Se denomina *afijación de Neyman*, debido a la contribución importante de [Neyman 1934](#).

Como se puede apreciar, el cálculo de los tamaños de muestra  $n_h$  requiere que las cuasidesviaciones típicas de cada subpoblación sean conocidas. Asimismo, en el caso del problema de minimizar la varianza dado un coste fijado  $C$ , la varianza mínima solo se podrá obtener si  $S_{yU_h}$  es conocida para todo  $h = 1, \dots, H$ . Normalmente en la práctica estos valores no están disponibles. Para solventar este problema se podrían usar aproximaciones cercanas a las cuasidesviaciones típicas verdaderas haciendo uso de la experiencia pasada en el caso de encuestas repetidas en el tiempo. La afijación obtenida de esta forma podría ser cercana a la óptima. Existen otras alternativas que se muestran a continuación, como usar información auxiliar correlada con la variable de estudio o aplicar la afijación proporcional.

#### 6.3.2 Afijación óptima con información auxiliar

Sea  $x$  una variable auxiliar que presenta una correlación alta con la variable de estudio  $y$ , tal que las cuasidesviaciones típicas poblacionales de  $x$ ,  $S_{xU_h}$ , son conocidas. Bajo estas condiciones, un método que suele ser usado en la práctica con buenos resultados es considerar la información conocida de esta variable auxiliar  $x$  en la expresión de la afijación óptima de la siguiente forma:

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum_{h=1}^H N_h \cdot S_{xU_h}}$$

Si la correlación entre  $x$  e  $y$  es perfecta, esto es,  $y_k = a + b \cdot x_k, k = 1, \dots, N$ , entonces la afijación es óptima, ya que en este caso  $S_{yU_h}^2 = b^2 \cdot S_{xU_h}^2$ . Por tanto:

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum_{h=1}^H N_h \cdot S_{xU_h}} = n \cdot \frac{N_h \cdot \frac{S_{yU_h}}{b}}{\sum_{h=1}^H N_h \cdot \frac{S_{yU_h}}{b}} = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^H N_h \cdot S_{yU_h}},$$

que es la afijación de Neyman. Si la correlación no es perfecta pero es fuerte, la expresión de  $n_h$  suele conducir a valores cercanos a la afijación óptima.

### 6.3.3 Afijación proporcional

La afijación proporcional puede ser una buena alternativa ya que si los tamaños poblacionales de los estratos  $N_h$  son conocidos, entonces este tipo de afijación siempre puede ser calculada para un tamaño de muestra fijado  $n$ . La afijación proporcional se define como

$$n_h = n \cdot \frac{N_h}{N} \quad (6.10)$$

Esto es, el número de unidades muestrales en cada estrato es proporcional al tamaño del estrato.

Si las cuasidesviaciones típicas  $S_{yU_h}$  son todas iguales, esto es,  $S_{yU_h} \equiv S$  entonces la afijación proporcional es óptima. En efecto:

$$n_h = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^H N_h \cdot S_{yU_h}} = n \cdot \frac{N_h \cdot S}{\sum_{h=1}^H N_h \cdot S} = n \cdot \frac{N_h}{N}$$

**Comentario 28.** Por definición, la probabilidad de que un elemento de la población aparezca en la muestra  $s$  es igual a la probabilidad de que dicho elemento aparezca en la muestra  $s_h$ :

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(k \in s_h) = \frac{n_h}{N_h}$$

Si se utiliza la afijación proporcional, entonces  $\pi_k = \frac{n}{N}$ , que es equivalente a la probabilidad de que el elemento  $k$  pertenezca a una muestra aleatoria simple,  $\forall k = 1, \dots, N$ . Sin embargo, una “mala” muestra, poco representativa de la población, por ejemplo con todos los elementos pertenecientes a un único estrato, no podría ocurrir en una muestra estratificada con afijación proporcional.

Asimismo, cada elemento de la muestra tiene el mismo *factor de expansión* (o *peso de muestreo*), por lo que todos los elementos representan el mismo número de unidades de la población. Cuando ocurre esto se dice que la muestra es *autoponderada*.

#### Definición 18

Un diseño de muestreo  $p(s)$  se dice que produce muestras *autoponderadas* respecto a un estimador del parámetro poblacional  $\theta$  si dicho estimador ( $\hat{\theta}$ ) se puede expresar en función del total muestral de la siguiente forma:

$$\hat{\theta} = K \sum_{k \in s} y_k$$

donde  $K$  es una constante y se denomina *factor de expansión*.

El muestreo aleatorio estratificado con afijación proporcional produce muestras *autoponderadas* respecto al estimador de Horvitz-Thompson. En efecto, para el caso del total poblacional, se tiene:

$$\hat{Y}_U^{\text{HT}} = \sum_{h=1}^H N_h \cdot \bar{y}_{s_h} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k \in s_h} y_k = \frac{N}{n} \sum_{h=1}^H \sum_{k \in s_h} y_k = \frac{N}{n} \sum_{k \in s} y_k$$

Como se puede observar, todas las observaciones aparecen multiplicadas por un mismo factor  $\frac{N}{n}$ .

■

### 6.3.4 Afijación proporcional al total de la variable $y$

Suponiendo que la variable  $y$  toma valores positivos, la afijación proporcional al total de la variable  $y$  se define como

$$n_h = n \cdot \frac{\sum_{k \in U_h} y_k}{\sum_{k \in U} y_k}$$

Para poder aplicar este tipo de afijación, los totales de la variable  $y$  en cada estrato deben ser conocidos y esto precisamente no suele ocurrir, por lo que en la práctica no podrá ser usada.

Por otra parte, se puede observar que esta afijación es óptima si el coeficiente de variación

$$cv(Y_{U_h}) = \frac{S_{yU_h}}{\bar{y}_{U_h}}$$

es constante en todos los estratos,  $cv(Y_{U_h}) \equiv cv, \forall h = 1, \dots, H$ .

$$\begin{aligned} n_h &= n \cdot \frac{\sum_{k \in U_h} y_k}{\sum_{k \in U} y_k} = n \cdot \frac{\sum_{k \in U_h} y_k}{\sum_{h=1}^H \sum_{k \in U_h} y_k} = n \cdot \frac{N_h \cdot \bar{y}_{U_h}}{\sum_{h=1}^H N_h \cdot \bar{y}_{U_h}} = \\ &= n \cdot \frac{N_h \cdot \frac{S_{yU_h}}{cv}}{\sum_{h=1}^H N_h \cdot \frac{S_{yU_h}}{cv}} = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^H N_h \cdot S_{yU_h}} \end{aligned}$$

### 6.3.5 Afijación proporcional al total de una variable auxiliar

Sea  $x$  una variable auxiliar que toma valores positivos y que presenta una correlación alta con la variable de estudio  $y$ , tal que los totales de la variable en cada estrato  $\sum_{k \in U_h} x_k$  son conocidos. Se define la afijación proporcional al total de la variable  $x$  como

$$n_h = n \cdot \frac{\sum_{k \in U_h} x_k}{\sum_{k \in U} x_k}$$



Este tipo de afijación ha resultado ser útil en la práctica. La justificación de su uso radica en que si  $x$  e  $y$  tienen correlación alta y el coeficiente de variación es aproximadamente el mismo en todos los estratos, entonces esta afijación no debería estar lejos de la óptima.

*Ejemplo 38.* En el contexto del Ejemplo 37 supongamos ahora que conocemos el total de la variable de estratificación  $x$ , que denota el número de escaños en los consejos municipales, para cada estrato. Por consiguiente, se dispone del total poblacional de dicha variable. La información viene recogida en la siguiente tabla:

Número de escaños	$N_h$	$\sum_{k \in U_h} x_k$	$\sum_{k \in U_h} x_k^2$
31 – 40	44	1518	52764
41 – 50	168	7524	339344
51 – 70	56	3198	184168
71 o más	16	1260	100016

Se desea obtener una muestra estratificada de 40 para estimar el total de escaños del Partido Socialdemócrata. Dado que no tenemos información sobre la variable de estudio, para determinar la distribución de la muestra entre los distintos estratos, se podría usar la afijación proporcional, la afijación proporcional al total de  $x$  o bien la afijación óptima con la información auxiliar de  $x$ .

Afijación proporcional:

$$n_h = n \cdot \frac{N_h}{N} \implies n_1 = 6, n_2 = 24, n_3 = 8, n_4 = 2$$

Afijación proporcional al total de la variable auxiliar  $x$ :

$$n_h = n \cdot \frac{\sum_{k \in U_h} x_k}{\sum_{k \in U} x_k} \implies n_1 = 5, n_2 = 22, n_3 = 9, n_4 = 4$$

Afijación óptima con información auxiliar ( $x$ ):

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum_{h=1}^H N_h \cdot S_{xU_h}} \implies n_1 = 5, n_2 = 21, n_3 = 10, n_4 = 4$$

■

## 6.4 Comparación de la precisión del estimador de Horvitz-Thompson en muestreo aleatorio estratificado según el tipo de afijación y el muestreo aleatorio simple

En primer lugar se va a realizar la comparación de la precisión del estimador HT bajo el diseño del muestreo aleatorio estratificado considerando la afijación óptima con

respecto a la afijación proporcional, para comprobar que el estimador es siempre igual o más preciso cuando se usa la afijación óptima:

$$\mathbb{V}_{strs,opt}(\hat{Y}_U^{HT}) \leq \mathbb{V}_{strs,prop}(\hat{Y}_U^{HT}) \quad (6.11)$$

donde  $\mathbb{V}_{strs,opt}(\hat{Y}_U^{HT})$  es la varianza bajo muestreo aleatorio estratificado con afijación óptima y  $\mathbb{V}_{strs,prop}(\hat{Y}_U^{HT})$  con afijación proporcional.

La varianza del estimador HT bajo muestreo aleatorio estratificado con afijación óptima es, usando la afijación de Neyman dada en (6.9):

$$\begin{aligned} \mathbb{V}_{strs,opt}(\hat{Y}_U^{HT}) &= \sum_{h=1}^H N_h^2 \cdot (1 - f_h) \cdot \frac{S_{yU_h}^2}{n_h} = \\ &= N^2 \sum_{h=1}^H \frac{W_h^2 \cdot S_{yU_h}^2}{n \cdot \frac{W_h \cdot S_{yU_h}}{\sum_{h=1}^H W_h \cdot S_{yU_h}}} - N \sum_{h=1}^H W_h \cdot S_{yU_h}^2 = \\ &= \frac{N^2}{n} \cdot \left( \sum_{h=1}^H W_h \cdot S_{yU_h} \right)^2 - N \sum_{h=1}^H W_h \cdot S_{yU_h}^2 \end{aligned}$$

La varianza bajo muestreo aleatorio estratificado con afijación proporcional es, usando la afijación dada en (6.10):

$$\begin{aligned} \mathbb{V}_{strs,prop}(\hat{Y}_U^{HT}) &= \sum_{h=1}^H N_h^2 \cdot (1 - f_h) \cdot \frac{S_{yU_h}^2}{n_h} = N^2 \sum_{h=1}^H \frac{W_h^2 \cdot S_{yU_h}^2}{n \cdot \frac{N_h}{N}} - N \sum_{h=1}^H W_h \cdot S_{yU_h}^2 = \\ &= N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \sum_{h=1}^H W_h \cdot S_{yU_h}^2 \end{aligned}$$

Restando las dos expresiones, se obtiene lo que se pretendía demostrar:

$$\begin{aligned} \mathbb{V}_{strs,prop}(\hat{Y}_U^{HT}) - \mathbb{V}_{strs,opt}(\hat{Y}_U^{HT}) &= \frac{N^2}{n} \sum_{h=1}^H W_h \cdot S_{yU_h}^2 - \frac{N^2}{n} \cdot \left( \sum_{h=1}^H W_h \cdot S_{yU_h} \right)^2 = \\ &= \frac{N^2}{n} \cdot \left[ \sum_{h=1}^H W_h \cdot S_{yU_h}^2 - \left( \sum_{h=1}^H W_h \cdot S_{yU_h} \right)^2 \right] = \\ &= \frac{N^2}{n} \sum_{h=1}^H W_h \cdot (S_{yU_h} - \bar{S}_y)^2 \geq 0 \end{aligned}$$

donde  $\bar{S}_y = \sum_{h=1}^H W_h \cdot S_{yU_h}$  es la media de las cuasidesviaciones típicas de los estratos.

**Comentario 29.** Obsérvese que la igualdad de precisiones se obtiene cuando las cuasi-desviaciones típicas  $S_{yU_h}$  son iguales en todos los estratos. La ganancia en precisión de la afijación óptima respecto de la proporcional será mayor cuanto más variación exista en las cuasivarianzas de los estratos. ■

Comparemos la precisión obtenida con el estimador HT bajo el diseño del muestreo aleatorio estratificado con afijación proporcional con respecto al diseño del muestreo aleatorio simple.

De la descomposición del análisis de la varianza, tenemos que la variación total,  $SST$ , se puede descomponer como la suma de la variación interestrato,  $SSB$ , y la variación intraestrato,  $SSW$ .

$$(N - 1) \cdot S_{yU}^2 = \sum_{k \in s} (y_k - \bar{y}_U)^2 = \sum_{h=1}^H N_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H (N_h - 1) \cdot S_{yU_h}^2 \quad (6.12)$$

De forma equivalente, podemos representarlo como

$$SST = SSB + SSW$$

En efecto:

$$\begin{aligned} SST &= \sum_{h=1}^H \sum_{k \in U_h} (y_k - \bar{y}_U)^2 = \sum_{h=1}^H \sum_{k \in U_h} (y_k - \bar{y}_{U_h} + \bar{y}_{U_h} - \bar{y}_U)^2 = \\ &= \sum_{h=1}^H \sum_{k \in U_h} (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2 + 2 \sum_{h=1}^H \sum_{k \in U_h} (y_k - \bar{y}_{U_h}) \cdot (\bar{y}_{U_h} - \bar{y}_U) = \\ &= \underbrace{\sum_{h=1}^H N_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2}_{SSB} + \underbrace{\sum_{h=1}^H (N_h - 1) \cdot S_{yU_h}^2}_{SSW} \end{aligned}$$

ya que el último sumando es 0.

Recordemos que la varianza del estimador HT bajo el diseño aleatorio simple sin reemplazamiento es

$$\begin{aligned} \mathbb{V}_{srswor}(\hat{Y}_U^{\text{HT}}) &= N^2 \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot S_{yU}^2 = \\ &= \frac{N^2}{N-1} \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot \left[ \sum_{h=1}^H N_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H (N_h - 1) \cdot S_{yU_h}^2 \right] = \\ &= \frac{N^3}{N-1} \cdot \left( \frac{1}{n} - \frac{1}{N} \right) \cdot \left[ \sum_{h=1}^H W_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H \left( W_h - \frac{1}{N} \right) \cdot S_{yU_h}^2 \right] \end{aligned}$$

donde se ha usado (6.12).

Por tanto:

$$\begin{aligned}
 \mathbb{V}_{srsor}(\hat{Y}_U^{\text{HT}}) - \mathbb{V}_{strs}(\hat{Y}_U^{\text{HT}}) &= \\
 &= \frac{N^3}{N-1} \left( \frac{1}{n} - \frac{1}{N} \right) \left[ \sum_{h=1}^H W_h (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H \left( W_h - \frac{1}{N} \right) S_{y_{U_h}}^2 - \frac{N-1}{N} \sum_{h=1}^H W_h S_{y_h}^2 \right] = \\
 &= \frac{N^3}{N-1} \left( \frac{1}{n} - \frac{1}{N} \right) \left[ \sum_{h=1}^H W_h (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H \left( W_h - \frac{1}{N} \right) S_{y_{U_h}}^2 - \frac{N-1}{N} \sum_{h=1}^H W_h S_{y_h}^2 \right] = \\
 &= \frac{N^3}{N-1} \left( \frac{1}{n} - \frac{1}{N} \right) \left[ \sum_{h=1}^H W_h (\bar{y}_{U_h} - \bar{y}_U)^2 - \frac{1}{N} \sum_{h=1}^H (1 - W_h) S_{y_{U_h}}^2 \right]
 \end{aligned}$$

**Comentario 30.** Del resultado anterior, se puede afirmar que es teóricamente posible que el muestreo aleatorio estratificado con afijación proporcional conduzca a una varianza ligeramente mayor que el diseño aleatorio simple sin reemplazamiento, en el caso de que todos los estratos presenten medias  $\bar{y}_{U_h}$  iguales o aproximadamente iguales. Sin embargo, en la mayoría de los casos esta igualdad no se da y normalmente  $\sum_{h=1}^H W_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2$  excederá considerablemente el valor de  $\frac{1}{N} \sum_{h=1}^H (1 - W_h) \cdot S_{y_{U_h}}^2$ . ■

**Comentario 31.** Si la variación interestrato representa una proporción significativa de la variación total, entonces las diferencias entre las medias de los estratos son una de las principales razones de la variación de  $y$ . En tal caso, el diseño *strs* con afijación proporcional produce una varianza sustancialmente menor. Así pues, cuanto más difieran las medias de los estratos más ganancia en precisión se obtendrá con la estratificación respecto a aplicar muestreo aleatorio simple. ■

*Ejemplo 39.* En el contexto del Ejemplo 37 y el Ejemplo 38, supongamos que se conoce el total del número de escaños en los consejos municipales, para cada estrato. Por consiguiente, se dispone del total poblacional de la variable de estudio  $y$ . La información viene recogida en la siguiente tabla:

Número de escaños	$N_h$	$\sum_{k \in U_h} y_k$	$\sum_{k \in U_h} y_k^2$
31 – 40	44	756	13784
41 – 50	168	3383	72223
51 – 70	56	1545	44529
71 o más	16	617	24137

Calculemos la varianza para las tres afijaciones obtenidas en el Ejemplo 38 y comparemos el resultado respecto a aplicar el diseño del muestreo aleatorio simple usando el estimador HT.

Las cuasivarianzas poblacionales de cada estrato son:

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \cdot \left( \sum_{k \in U_h} y_k^2 - N_h \cdot \bar{y}_{U_h}^2 \right)$$

$$S_{yU_1}^2 \approx 18,48; \quad S_{yU_2}^2 \approx 24,55; \quad S_{yU_3}^2 \approx 34,61; \quad S_{yU_4}^2 \approx 22,93$$

La varianza del estimador usando la afijación proporcional es

$$\mathbb{V}_{strs,prop}(\hat{Y}_U^{HT}) = \sum_{h=1}^H N_h^2 \cdot \frac{1 - f_h}{n_h} \cdot S_{yU_h}^2 \approx 44092,60$$

La varianza es ligeramente menor aplicando muestreo estratificado con afijación proporcional, ya que la varianza obtenida con afijación proporcional al total de  $x$  y la varianza aplicando afijación óptima con información auxiliar es, respectivamente:

$$\mathbb{V}_{strs,prop_x} \approx 44934,73; \quad \mathbb{V}_{strs,opt_x} \approx 45228,54$$

Para calcular la varianza del estimador de expansión bajo muestreo aleatorio simple, debemos obtener en primer lugar la cuasivarianza poblacional de  $y$ .

$$S_{yU}^2 = \frac{1}{N - 1} \cdot \left[ \sum_{h=1}^H N_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^H (N_h - 1) \cdot S_{yU_h}^2 \right] \approx 52,56$$

donde se ha usado  $\bar{y}_{U_1} = \frac{189}{11}$ ,  $\bar{y}_{U_2} = \frac{3383}{168}$ ,  $\bar{y}_{U_3} = \frac{1545}{56}$ ,  $\bar{y}_{U_4} = 38,5625$ .

Se sabe que la variable de estratificación  $x$  tiene correlación positiva alta con la variable de estudio  $y$  (es aproximadamente 0,76), por lo que cabe esperar que la varianza obtenida bajo muestreo aleatorio simple con tamaño  $n = 40$  y usando el estimador de expansión HT sea mucho mayor que usando muestreo estratificado con las afijaciones que tienen en cuenta la información de  $x$ . Asimismo, dado que  $SSB$  representa un porcentaje significativo de la variación total (aproximadamente 52%), se podría esperar una varianza sustancialmente menor que con el diseño aleatorio simple sin reemplazamiento.

En efecto, la varianza del diseño aleatorio simple sin reemplazamiento es

$$\mathbb{V}(\hat{Y}_U^{HT}) = N^2 \cdot \frac{1 - f}{n} \cdot S_{yU}^2 \approx 91058,80$$

Se observa una ganancia importante con la estratificación, tal y como cabía esperar.

■

## Bibliografía

- Cochran, W.G. (1977). *Sampling Techniques*. 3rd. New York: Wiley.
- Danielsson, S. (1975). "Optimal allokering vid vissa klasser av urvalsförfaranden". Tesis doct. Department of Mathematics, University of Linköping, Sweden.
- Neyman, J. (1934). "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection". En: *Journal of the Royal Statistical Society* 97, págs. 558-625.
- Rao, J. N. K. (1979). "Optimization in the design of sample surveys". En: J. S. Rustagi (ed.), *Optimizing Methods in Statistics: Proceedings of an International Conference*. New York: Academic Press, págs. 419-434.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
- Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. 3rd. London: Charles Griffin y Co.

## Tema 7

### **Estimación insesgada en diseños muestrales por conglomerados I. Muestreo por conglomerados sin submuestreo: definición, estimadores, varianza y estimador de la varianza.**

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

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

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.

### **7.1 Muestreo por conglomerados sin submuestreo: definición, estimadores, varianza y estimador de la varianza**

#### **7.1.1 Introducción y definiciones**

Los diseños discutidos en los Temas 3 a 6 de 'Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas' asumen que es posible el muestreo directo de los elementos. Es decir, los elementos de la población se pueden usar como elementos muestrales en una única etapa de muestreo. Sin embargo, en muchas encuestas de tamaño mediano y grande, el muestreo directo de elementos no se usa por una o por las dos razones siguientes:

- i. No existe un marco muestral que identifique a todos y cada uno de los elementos de la población, y la producción de tal marco puede ser difícil, cara o imposible.

Por ejemplo, no se puede construir un listado con todos los clientes de una tienda, pero sí podemos construir una lista de todos los individuos de una ciudad para los cuales sólo existe una lista de viviendas; sin embargo, la construcción de esta lista consumirá mucho tiempo y será muy cara.

- ii. Los elementos de la población están dispersos sobre un área muy grande, en cuyo caso el muestreo directo de elementos dará lugar a una muestra demasiado dispersa. Por tanto, el coste del trabajo de campo sería prohibitivo debido al alto coste de los viajes si son necesarias entrevistas personales. También puede ser dificultosa la supervisión del trabajo de campo, lo que puede dar lugar a una tasa de falta de respuesta muy grande y a errores de muestreo graves.

Si la población objetivo son los alumnos de un determinado curso, por ejemplo cuarto de ESO, en España, es más barato tomar una muestra de colegios y entrevistar a todos los alumnos de cuarto de los colegios seleccionados que entrevistar a toda una muestra aleatoria simple de alumnos de cuarto de ESO. Con una muestra aleatoria simple de alumnos es posible que en algún caso hubiera que viajar hasta algún colegio para entrevistar a un único alumno.

*Ejemplo 40.* [Basow y Silberg 1987](#) escribieron un artículo con los resultados que obtuvieron de una investigación que realizaron sobre si los estudiantes evalúan de manera distinta a sus profesores universitarios que a sus profesoras universitarias. Los autores formaron parejas de 16 profesoras y 16 profesores por materia, años de experiencia docente y nivel de conocimientos. A continuación proporcionaron cuestionarios a los alumnos que pertenecían a los grupos de estos profesores. El tamaño de muestra para realizar este estudio es  $n = 32$ , el número de académicos analizados, no es 1029, el número de alumnos que rellenaron los cuestionarios.

Las evaluaciones de los alumnos reflejan los distintos estilos de enseñanza de los profesores. Es probable que los alumnos del mismo grupo coincidan en la evaluación que hicieron del profesor y no deban, entonces, considerarse como observaciones independientes, pues es probable que sus calificaciones estén correlacionadas en forma positiva. Si se ignora esta correlación positiva y las calificaciones se tratan como observaciones independientes, las diferencias se pueden considerar estadísticamente significativas con una mayor frecuencia de la debida. ■

Hay una gran variedad de diseños de muestreo para las encuestas en las cuales el muestreo directo de elementos es imposible o no es práctico. Varían desde el muestreo por conglomerados a diseños muy complejos de muestreo en varias etapas usando probabilidades de muestreo distintas en cada etapa de muestreo<sup>1</sup>. Veamos en primer lugar varios términos importantes. En *muestreo por conglomerados*, la población finita está agrupada en subpoblaciones llamadas *conglomerados*. Se selecciona una probabilidad de muestreo del conglomerado y todos los elementos de la población del conglomerado seleccionado son encuestados.

---

<sup>1</sup>Los diseños de muestreo complejos se explican en 'Producción Estadística Oficial: Métodos Avanzados de la especialidad de Estadística-Ciencia de datos'



El muestreo por conglomerados también se llama *muestreo por conglomerados en una única etapa*. Por contra, en el *muestreo de conglomerados con submuestreo* la muestra de elementos se obtiene como resultado de dos etapas de muestreo:

- i. Los elementos de la población se agrupan en primer lugar en subpoblaciones disjuntas, llamadas *unidades muestrales primarias* (PSUs del inglés *primary sampling units*). Se obtiene la probabilidad de selección de cada PSU (*muestreo de primera etapa*).
- ii. Para cada PSU de primera etapa, se decide el tipo de unidad de muestreo a usar en el muestreo de la segunda etapa. Estas *unidades muestrales de segunda etapa* (SSUs del inglés *secondary sampling units*) pueden ser elementos o conglomerados de elementos. Se obtiene la probabilidad muestral de las SSUs para cada PSU en el muestreo de primera etapa. Cuando las SSUs son conglomerados, se encuesta a cada unidad en las SSUs seleccionados en el caso de utilizar solo dos etapas; en caso contrario, se sigue muestreando.

**Comentario 32.** Cuando cada SSU es un elemento, usamos el término muestreo bietápico; cuando cada SSU es un conglomerado de elemento, usamos el término muestreo de conglomerados con submuestreo o muestreo multietápico. ■

*Ejemplo 41.* La Encuesta sobre Uso de Drogas en Enseñanzas Secundarias en España, ESTUDES<sup>2</sup>, es una operación estadística que tiene por objetivo recabar información de valor para diseñar y evaluar políticas dirigidas a prevenir el consumo de drogas y otras adicciones y los problemas derivados del mismo. Para ello se realiza una encuesta a estudiantes de 14 a 18 años matriculados en la ESO, Bachillerato, Ciclos de Formación Profesional Básica y Ciclos Formativos de Grado Medio de Formación Profesional.

Para ello se realiza un muestreo por conglomerados bietápico, en el que, en primera etapa, se seleccionan aleatoriamente centros educativos (unidades de primera etapa) y, en segundo lugar, aulas (unidades de segunda etapa), cumplimentando el cuestionario a todos los alumnos presentes en las mismas. En este caso los centros educativos son las PSUs y las aulas son las SSUs. Los alumnos son los elementos.

Supongamos que en lugar de seleccionar aulas en la segunda etapa se encuestase a todos los alumnos de los centros seleccionados en la primera etapa. En tal caso se trataría de un muestreo por conglomerados en una etapa, las PSUs serían los centros educativos y no habría SSUs. ■

---

<sup>2</sup>[https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/sistemaInformacion/encuestas\\_ESTUDES.htm](https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/sistemaInformacion/encuestas_ESTUDES.htm)

**Definición 19**

El *muestreo multietápico* consiste en un muestreo en tres o más etapas. Hay una jerarquía de unidades muestrales: las unidades muestrales de primera etapa, las unidades muestrales secundarias dentro de las PSUs, las unidades muestrales terciarias dentro de las SSUs, y así sucesivamente. Las unidades muestrales en la última etapa de muestreo se llaman *unidades de muestreo últimas* y las de aquellas en la etapa anterior a la última se llaman *unidades de muestreo penúltimas*.

**Comentario 33.** El término muestreo multietápico de elementos se aplica cuando las unidades de muestreo últimas son elementos. En el muestreo multietápico de conglomerados, las unidades de muestreo últimas son conglomerados de elementos. ■

*Ejemplo 42.* La Encuesta Europea de Salud en España<sup>3</sup> es una operación estadística dirigida al conjunto de personas de 15 y más años que reside en viviendas familiares en todo el territorio nacional. Su objetivo principal es obtener datos sobre el estado de salud, la utilización de los servicios sanitarios y los factores determinantes de salud, de manera armonizada y comparable a nivel europeo.

Para ello utiliza un muestreo trietápico. Las unidades de primera etapa son las secciones censales. Las unidades de segunda etapa son las viviendas familiares principales, investigándose a todos los hogares que tienen su residencia habitual en las mismas. Dentro de cada hogar se selecciona a un adulto (15 o más años).

Esto es un ejemplo de muestreo trietápico en el que las secciones censales son las PSUs, las viviendas son las SSUs y el adulto que rellena el cuestionario es la unidad muestral terciaria o unidad muestral final. ■

En este tema nos concentraremos en la estimación del total poblacional  $Y_U = \sum_U y_k$ . Cabe señalar que con este diseño la estimación de la media poblacional  $\bar{y}_U = \frac{Y_U}{N}$  no se puede obtener como se hace habitualmente dividiendo  $\hat{Y}_U^{\text{HT}}$  por  $N$  y  $\mathbb{V}(\hat{Y}_U^{\text{HT}})$  y  $\hat{\mathbb{V}}(\hat{Y}_U^{\text{HT}})$  por  $N^2$ . Esta complicación surge porque el tamaño poblacional  $N$  normalmente es desconocida en las encuestas que necesitan usar el muestreo por conglomerados. Siendo  $N$  desconocido, el parámetro  $\bar{y}_U = \frac{Y_U}{N}$  es un cociente de dos parámetros desconocidos, para lo que es necesario estimadores separados.

### 7.1.2 Estimadores, varianza y estimador de la varianza

En el muestreo por conglomerados, la población finita  $U = \{1, \dots, k, \dots, N\}$  se divide en  $N_I$  subpoblaciones, llamados conglomerados, y denotados por  $U_1, \dots, U_i, \dots, U_{N_I}$ . El

<sup>3</sup>[https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176784&menu=metodologia&idp=1254735573175](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176784&menu=metodologia&idp=1254735573175)

conjunto de conglomerados se representa simbólicamente por

$$U_I = \{1, \dots, i, \dots, N_I\}.$$

Esto representa una población de conglomerados de la que se selecciona una muestra de conglomerados. El índice  $I$  se utilizará para identificar entidades asociadas con los conglomerados. La razón para usar  $I$  en lugar de  $C$ , que sería más natural, es que el primero facilita la transición al muestreo de conglomerados con submuestreo (véase el tema 1<sup>4</sup> del bloque “Producción Estadística Oficial: Métodos Avanzados” de la especialidad de Estadística y Ciencia de datos del grupo B de materias específicas), en que  $I$  se referirá a la primera etapa,  $II$  a la segunda etapa y si hubiese más etapas se podrían seguir usando estos subíndices.

El número de elementos de la población en el  $i$ -ésimo conglomerado  $U_i$  se denota por  $N_i$ . La partición de  $U$  se expresa mediante las ecuaciones

$$U = \bigcup_{i \in U_I} U_i \quad \text{y} \quad N = \sum_{i \in U_I} N_i.$$

### Definición 20

El *muestreo por conglomerados en una única etapa* (o simplemente *muestreo por conglomerados*) se define ahora de la siguiente forma:

- i. Una muestra probabilística  $s_I$  de conglomerados se selecciona a partir de  $U_I$  de acuerdo con el diseño  $p_I(\cdot)$ . El tamaño de  $s_I$  se denota por  $n_I$ , para un diseño de tamaño fijo, o por  $n_{s_I}$  para un diseño de tamaño variable.
- ii. Cada elemento poblacional de los conglomerados seleccionados es observado (entrevistado).

Aquí  $p_I(\cdot)$  puede ser cualquiera de los diseños convencionales, es decir, muestreo aleatorio simple sin reemplazamiento, muestreo sistemático, muestreo estratificado, etcétera. Seguiremos usando  $s$  como símbolo del conjunto de elementos que se observan. Es decir,

$$s = \bigcup_{i \in s_I} U_i.$$

El tamaño de  $s$  es

$$n_s = \sum_{i \in s_I} N_i.$$

Señalamos que incluso si  $p_I(\cdot)$  es un diseño de tamaño fijo, el número de elementos observados  $n_s$  en general no será fijo, porque los tamaños de los conglomerados  $N_i$

<sup>4</sup>Tema 1. Estimación insesgada en diseños muestrales por conglomerados II. Muestreo de conglomerados con submuestreo: definición, estimadores, varianza y estimador de la varianza.

pueden variar.

Las probabilidades de inclusión del conglomerado de primer y segundo orden inducidas por el diseño  $p_I(\cdot)$  son

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I)$$

y para dos conglomerados  $i$  y  $j$

$$\pi_{Iij} = \sum_{s_I \ni i, j} p_I(s_I).$$

Se verifica que  $\pi_{Iii} = \pi_{Ii}$ . Volvamos a las probabilidades de inclusión del elemento. Como la muestra  $s$  contiene todos los elementos en los conglomerados seleccionados, para cada  $k$  en  $U_i$ , tenemos

$$\pi_k = \mathbb{P}(s \ni k) = \mathbb{P}(s_I \ni i) = \pi_{Ii}. \quad (7.1)$$

Las probabilidades de inclusión de segundo orden vienen dadas por

$$\pi_{kl} = \mathbb{P}(s \ni k, l) = \mathbb{P}(s_I \ni i) = \pi_{Ii} \quad (7.2)$$

si tanto  $k$  como  $l$  pertenecen al mismo conglomerado  $U_i$ , y

$$\pi_{kl} = \mathbb{P}(s \ni k, l) = \mathbb{P}(s_I \ni i, j) = \pi_{Iij} \quad (7.3)$$

si  $k$  y  $l$  pertenecen a distintos conglomerados  $U_i$  y  $U_j$ . Cabe señalar que  $\pi_{kk} = \pi_k$ . Es conveniente introducir la siguiente notación

$$Y_{U_i} = \sum_{k \in U_i} y_k$$

para el total del conglomerado  $i$ -ésimo. El total poblacional a estimar se puede expresar entonces como

$$Y_U = \sum_{k \in U} y_k = \sum_{i \in U_I} Y_{U_i}.$$

Sea  $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ . Tenemos entonces el siguiente resultado:

### Teorema 23

En el muestreo por conglomerados, el estimador de Horvitz-Thompson del total poblacional  $Y_U = \sum_{k \in U} y_k$  se puede escribir como

$$\hat{Y}_U^{\text{HT}} = \sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}}. \quad (7.4)$$

La varianza viene dada por

$$\mathbb{V}[\hat{Y}_U^{\text{HT}}] = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}}. \quad (7.5)$$

Un estimador insesgado de la varianza es

$$\hat{\mathbb{V}}[\hat{Y}_U^{\text{HT}}] = \sum_{i \in s_I} \sum_{j \in s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}}. \quad (7.6)$$

### Demostración 23

Para determinar el estimador de Horvitz-Thompson,

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{i \in s_I} \sum_{k \in U_i} \frac{y_k}{\pi_k}$$

usamos la ecuación (7.1) y obtenemos

$$\hat{Y}_U^{\text{HT}} = \sum_{i \in s_I} \frac{(\sum_{k \in U_i} y_k)}{\pi_{Ii}} = \sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}}.$$

Los otros resultados se obtienen inmediatamente a partir de las propiedades del estimador de Horvitz-Thompson teniendo en cuenta que  $\sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}}$  es el estimador de Horvitz-Thompson de  $\sum_{i \in U_I} Y_{U_i}$ . Las probabilidades  $\pi$ s adecuadas son las probabilidades de inclusión de los conglomerados.

Si  $p_I(\cdot)$  es un diseño muestral de tamaño fijo, la varianza  $\mathbb{V}[\hat{Y}_U^{\text{HT}}]$  en el teorema 23 también se puede expresar como

$$\mathbb{V}[\hat{Y}_U^{\text{HT}}] = -\frac{1}{2} \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \left( \frac{Y_{U_i}}{\pi_{Ii}} - \frac{Y_{U_j}}{\pi_{Ij}} \right)^2 \quad (7.7)$$

con el estimador insesgado de la varianza

$$\hat{\mathbb{V}}(\hat{Y}_U^{\text{HT}}) = -\frac{1}{2} \sum_{i \in s_I} \sum_{j \in s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \left( \frac{Y_{U_i}}{\pi_{Ii}} - \frac{Y_{U_j}}{\pi_{Ij}} \right)^2. \quad (7.8)$$

El teorema 23 conduce a varias conclusiones interesantes sobre la eficiencia del muestreo por conglomerados. Asumimos que se usa el estimador de Horvitz-Thompson y que  $p_I(\cdot)$  es un diseño de tamaño fijo, por eso se utilizan las ecuaciones (7.7) y (7.8).

- A partir de (7.7), podemos ver que si todos los  $\check{Y}_{U_i} = \frac{Y_{U_i}}{\pi_{Ii}}$  son iguales, entonces  $\mathbb{V}[\hat{Y}_U^{\text{HT}}] = 0$ . Por tanto, si podemos elegir  $\pi_{Ii}$  aproximadamente proporcional a los

totales de los conglomerados  $Y_{U_i}$ , entonces el muestreo por conglomerados será altamente eficiente.

- b. Si los tamaños de los conglomerados  $N_i$  son conocidos en la fase de planificación, se puede elegir un diseño con  $\pi_{I_i} \propto N_i$ . Como  $Y_{U_i} = N_i \bar{y}_{U_i} = \sum_{k \in U_i} y_k$ , es una buena elección si hay poca variación entre las medias de los conglomerados  $\bar{y}_{U_i}$ . Si todos los  $\bar{y}_{U_i}$  son iguales, tendríamos, en efecto,  $\mathbb{V}[\hat{Y}_U^{\text{HT}}] = 0$ .
- c. Un diseño de muestreo por conglomerados con probabilidades iguales (es decir, uno en el que todos los  $\pi_{I_i}$  son iguales) a menudo es una mala opción cuando los conglomerados tienen distinto tamaño. Para que un diseño así sea eficiente, debemos tener  $\bar{y}_{U_i}$  aproximadamente proporcionales a  $N_i^{-1}$ . Esto, sin embargo, raramente se da en la práctica.

#### **Comentario 34. El muestreo por conglomerados y el muestreo estratificado.**

Los conglomerados recuerdan a los estratos, pero sólo de manera superficial: un conglomerado, al igual que un estrato, es una agrupación de los elementos de una población (individuos de una ciudad, empresas de una comunidad autónoma, establecimientos industriales de una provincia). Sin embargo, el proceso de construcción y selección es un poco distinto en estos dos métodos.

Mientras que, por lo general, la estratificación aumenta la precisión en relación con el muestreo aleatorio simple, el muestreo por conglomerados, con frecuencia, la disminuye. Los miembros de un mismo conglomerado tienden a ser más similares entre sí que los elementos seleccionados al azar de entre toda la población: los alumnos de cuarto de ESO de un mismo colegio tienden a tener un nivel de vida parecido; los peces de un mismo lago tienden a presentar concentraciones similares de mercurio. Por lo general, estas analogías surgen debido a ciertos factores subyacentes que podrían medirse o no. Por tanto, si extraemos una muestra con dos alumnos de un mismo colegio puede que no consigamos tanta información acerca de los alumnos de cuarto de ESO como la que obtendríamos al extraer una muestra de dos alumnos de colegios distintos. Al obtener una muestra de todos los elementos del mismo conglomerado, repetimos parcialmente la misma información en lugar de conseguir información nueva y esto implica una menor precisión para las estimaciones de las variables objetivo de la población.

El muestreo por conglomerados se utiliza en la práctica (sobre todo en las encuestas sociales) debido a que es más económico y conveniente obtener muestras por conglomerados que al azar entre la población de personas. Casi todas las encuestas sociales (a hogares, a individuos, a viviendas) utilizan el muestreo por conglomerados debido al ahorro de costes.

Un error que se suele cometer es usar encuestas que tienen un muestreo por conglomerados como si el muestreo fuese aleatorio simple. Esto puede provocar que los errores estándar obtenidos sean mucho menores de lo debido; lo que da la impresión de que los resultados de la encuesta son mucho más precisos de lo que realmente son. ■

### Comentario 35. Conglomerados de igual tamaño y conglomerados de distinto tamaño.

Dentro del muestreo por conglomerados se puede tener dos casos, aquel en el que cada conglomerado tiene el mismo número de elementos y aquel en el que los conglomerados tienen distintos tamaños.

En el primer caso, con todos los conglomerados de igual tamaño, que denotaremos por  $N_i = M$  para todo  $i$ . Los conglomerados de individuos en la práctica no se ajustan a este caso, pero podría aparecer en el muestreo agrícola. Es el caso más sencillo. Se usan los resultados del muestreo aleatorio simple con los totales por conglomerado como observaciones. El inconveniente es que casi siempre el muestreo por conglomerados con igual tamaño proporciona una menor precisión para los estimadores que en el caso de una muestra aleatoria simple con el mismo número de elementos.

Sin embargo, en las encuestas sociales es raro que los conglomerados tengan todos el mismo tamaño. La diferencia entre conglomerados con el mismo o distinto tamaño es que es probable que la variación entre los totales de los conglomerados  $Y_{U_i}$  individuales sea grande cuando los conglomerados tengan distinto tamaño. Así, esperaríamos que  $Y_{U_i}$  sea grande cuando el tamaño del conglomerado  $N_i$  fuese grande y que sea pequeño cuando  $N_i$  también lo sea. Con frecuencia,  $\widehat{V}[\widehat{Y}_U^{\text{HT}}]$  es mayor en una muestra por conglomerados cuando las PSUs tienen distintos tamaños que cuando todas las PSUs tienen el mismo número de SSUs.

En el caso de conglomerados de distinto tamaño el estimador del total poblacional no plantea problemas, pero sí el de la media poblacional. Si queremos obtener un estimador insesgado de la media poblacional y su varianza, para el caso de conglomerados de distinto tamaño, definimos  $K = \sum_{U_I} N_i$ , que es la cantidad total de SSUs en la población, y se tendría que  $\widehat{Y}_U^{\text{HT}} = \frac{\widehat{Y}_U^{\text{HT}}}{K}$ . Esto plantea el problema de que debemos conocer  $K$ , pero con frecuencia sólo conocemos el tamaño de los conglomerados que están en la muestra y no del resto, ya que como se vio al principio el muestreo por conglomerados se utiliza cuando no se dispone de marco poblacional o es muy difícil o costoso obtenerlo. ■

### 7.1.3 Muestreo por conglomerados aleatorio simple

En este apartado veremos un mayor desarrollo del punto (c) anterior: el muestreo por conglomerados con probabilidades iguales puede ser ineficiente a menudo. Consideremos un muestreo por conglomerados aleatorio simple sin reemplazamiento. Denotaremos este diseño por  $1st^5$ . Es decir, una muestra aleatoria simple sin reemplazamiento  $s_I$  de tamaño fijo  $n_I$  se selecciona de  $N_I$  conglomerados en  $U_I$  y todos los elementos en los conglomerados seleccionados son observados. Se sigue del Teorema 23 y de los resultados sobre el diseño muestral aleatorio simple sin reemplazamiento que el estimador de Horvitz-Thompson del total poblacional viene dado por

$$\widehat{Y}_U^{\text{HT}} = N_I \bar{y}_{s_I}, \quad (7.9)$$

<sup>5</sup>Del inglés *simple cluster random sampling without replacement*.

donde  $\bar{y}_{s_I} = \frac{1}{n_I} \sum_{i \in s_I} Y_{U_i}$  es la media de los totales del conglomerado  $Y_{U_i}$  en  $s_I$ . La varianza se puede escribir como

$$\mathbb{V}_{1st}(\hat{Y}_U^{\text{HT}}) = N_I^2 \frac{1 - f_I}{n_I} S_{Y_{U_I}}^2, \quad (7.10)$$

donde  $f_I = \frac{n_I}{N_I}$  es la fracción de muestreo de conglomerados y

$$S_{Y_{U_I}}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (Y_{U_i} - \bar{y}_{U_I})^2, \quad (7.11)$$

con  $\bar{y}_{U_I} = \sum_{i \in U_I} \frac{Y_{U_i}}{N_I}$ . El estimador insesgado de la varianza es

$$\hat{\mathbb{V}}_{1st}(\hat{Y}_U^{\text{HT}}) = N_I^2 \frac{1 - f_I}{n_I} S_{Y_{s_I}}^2, \quad (7.12)$$

donde

$$S_{Y_{s_I}}^2 = \frac{1}{n_I - 1} \sum_{i \in s_I} (Y_{U_i} - \bar{y}_{s_I})^2.$$

Para estimar  $\bar{y}_U$  en el caso de que todos los conglomerados sean de igual tamaño  $N_i = M$ , dividimos el total estimado entre el número de elementos, con lo que obtenemos:

$$\hat{\bar{Y}}_U^{\text{HT}} = \frac{\hat{Y}_U^{\text{HT}}}{N_I M}. \quad (7.13)$$

La varianza viene dada por

$$\mathbb{V}_{1st}(\hat{\bar{Y}}_U^{\text{HT}}) = \frac{1 - f_I}{n_I} \frac{S_{Y_{U_I}}^2}{M^2}. \quad (7.14)$$

Un estimador insesgado de la varianza es

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_U^{\text{HT}}) = \frac{1 - f_I}{n_I} \frac{S_{Y_{s_I}}^2}{M^2}. \quad (7.15)$$

*Ejemplo 43.* Un estudiante quiere estimar las calificaciones promedio de sus compañeros de residencia. En vez de obtener una lista de todos los alumnos que están en su residencia y realizar una muestra aleatoria simple, observa que dicha residencia consta de 100 dormitorios, cada uno con cuatro estudiantes. Elige cinco dormitorios al azar y pregunta a cada persona por sus calificaciones. Los resultados son los siguientes:

Número de persona	Dormitorio (Conglomerado)				
	1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08



Las unidades primarias son los dormitorios, de modo que  $N_I = 100$ ,  $n_I = 5$  y  $M = 4$ . La estimación del total de la población (la suma estimada de todas las calificaciones de todos los alumnos que pertenecen a la residencia, cantidad sin sentido en este ejemplo, pero útil para demostrar el procedimiento) es:

$$\hat{Y}_U^{\text{HT}} = \frac{100}{5}(12,16 + 11,36 + 8,96 + 12,96 + 11,08) = 1130,4$$

y

$$s_{Y_{s_I}}^2 = \frac{1}{5-1}[(12,16 - 11,304)^2 + \dots + (11,08 - 11,304)^2] = 2,256.$$

En este ejemplo,  $s_{Y_{s_I}}^2$  es solo la (cuasi)varianza muestral de los totales de los 5 dormitorios. Por tanto, usando (7.13) y (7.15) obtenemos

$$\hat{Y}_U^{\text{HT}} = \frac{1130,4}{400} = 2,826$$

y

$$\hat{V}\left(\hat{Y}_U^{\text{HT}}\right) = \frac{1 - \frac{5}{100}}{5} \frac{2,256}{4^2} = 0,027.$$

Obsérvese que en estos cálculos sólo se utiliza los valores de la última fila 'Total' de la tabla de datos, las calificaciones individuales se emplean sólo para calcular el total de cada dormitorio. ■

**Comentario 36.** Tal y como se ha señalado en el Tema 4 de este mismo bloque, el muestreo sistemático con un arranque aleatorio se corresponde formalmente con un diseño muestral 1st con  $n_I = 1$ , donde los  $N_I$  conglomerados se corresponden con las  $a$  posibles muestras sistemáticas. Además, el muestreo sistemático con  $m$  arranques aleatorios (véase la Sección 4.4) se puede considerar como un diseño muestral 1st con  $n_I = m$  y  $N_I = ma$ . En este caso, la ecuación (7.12) proporciona un estimador insesgado de la varianza para el estimador de Horvitz-Thompson del total poblacional. ■

Analizamos ahora más detenidamente el diseño 1st y lo comparamos con una selección directa aleatoria simple de elementos. Es útil trabajar con el coeficiente de homogeneidad  $\delta$ , definido en el contexto de este tema como

$$\delta = 1 - \frac{S_{yW}^2}{S_{yU}^2}, \quad (7.16)$$

donde

$$S_{yW}^2 = \frac{1}{N - N_I} \sum_{i \in U_I} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2 \quad (7.17)$$

es la varianza acumulada/agrupada dentro del conglomerado (*pooled within-cluster variance*, en inglés). Aquí,  $\bar{y}_{U_i} = \sum_{k \in U_i} \frac{y_k}{N_i}$  es la media del conglomerado  $i$ -ésimo. Cabe señalar que si

$$S_{yU_i}^2 = \frac{1}{N_i - 1} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$$

denota la varianza de  $y$  dentro del conglomerado  $U_i$ , podemos escribir la ecuación (7.17) como

$$S_{yW}^2 = \frac{\sum_{i \in U_I} (N_i - 1) S_{yU_i}^2}{\sum_{i \in U_I} (N_i - 1)}. \quad (7.18)$$

Esto proporciona una interpretación sencilla de  $S_{yW}^2$ . Es la media ponderada de las  $N_I$  varianzas de los conglomerados  $S_{yU_i}^2$ , con los pesos  $N_i - 1$ .

Como consecuencia,  $\delta$  es positivo o negativo según la media de la varianza dentro del conglomerado sea menor o mayor que la varianza total  $S_{yU}^2$ , respectivamente.

**Comentario 37.** El lector familiarizado con el análisis de regresión identificará  $\delta$  como el coeficiente de determinación ajustado por los grados de libertad, que a menudo se denota por  $R_{adj}^2$ , cuando se ajusta la regresión lineal de  $y$  sobre  $N_I$  variables dummy (indicando la pertenencia al conglomerado) de la población entera de  $N$  datos. ■

El coeficiente de homogeneidad  $\delta$  satisface

$$-\frac{N_I - 1}{N - N_I} \leq \delta \leq 1. \quad (7.19)$$

La cota superior de  $\delta$  se obtiene a partir de la ecuación (7.16) y del hecho que  $S_{yW}^2 \geq 0$ . Para verificar la cota inferior de  $\delta$ , podemos utilizar la descomposición estándar ANOVA  $SST = SSW + SSB$ , que en este caso particular es

$$(N - 1)S_{yU}^2 = (N - N_I)S_{yW}^2 + SSB \quad (7.20)$$

donde

$$SSB = \sum_{i \in U_I} N_i (\bar{y}_{U_i} - \bar{y}_U)^2.$$

Como  $SSB \geq 0$ , se sigue de la ecuación (7.20) que

$$\frac{S_{yW}^2}{S_{yU}^2} \leq \frac{N - 1}{N - N_I}.$$

Por tanto,

$$\delta = 1 - \frac{S_{yW}^2}{S_{yU}^2} \geq 1 - \frac{N - 1}{N - N_I} = -\frac{N_I - 1}{N - N_I}.$$

**Comentario 38.** Un valor pequeño de  $\delta$  significa que los elementos en el mismo conglomerado son diferentes con respecto a la variable de estudio, es decir, tienen un bajo grado de homogeneidad. Un valor grande de  $\delta$  significa que los elementos en el mismo conglomerado son similares, es decir, tienen un alto grado de homogeneidad.

En uno de los extremos,  $\delta = 1$  significa que la variación es cero dentro de cada conglomerado; en el otro extremo,  $\delta = -\frac{\bar{N}-1}{N-\bar{N}_I}$  implica que todas las medias de los conglomerados son iguales. El extremo inferior  $-\frac{\bar{N}-1}{N-\bar{N}_I}$  normalmente es cercano a cero, especialmente si  $N$  es grande comparado con  $N_I$ . El valor  $\delta = 0$  se obtiene cuando la varianza media dentro del conglomerado es igual a la varianza en la población entera  $U$ . ■

Denotemos por  $\bar{N} = \frac{N}{N_I}$  el número medio de elementos por conglomerado, sea  $K_I = \frac{N_I^2(1-f_I)}{n_I}$ , y sea

$$Cov = \frac{1}{N_I - 1} \sum_{i \in U_I} (N_i - \bar{N}) N_i \bar{y}_{U_i}^2 \quad (7.21)$$

la covarianza entre  $N_i$  y  $N_i \bar{y}_{U_i}^2$ . Se verifica fácilmente que

$$S_{YU}^2 = \bar{N} S_{yU}^2 \left( 1 + \frac{N - N_I}{N_I - 1} \delta \right) + Cov \quad (7.22)$$

Si introducimos esta expresión en (7.10), que denominamos  $V_{1st}$ , nos da

$$V_{1st} = \left( 1 + \frac{N - N_I}{N_I - 1} \delta \right) \bar{N} K_I S_{yU}^2 + K_I Cov. \quad (7.23)$$

El número esperado de elementos observados bajo 1st, con  $n_I$  conglomerados seleccionados a partir de  $N_I$ , es

$$\mathbb{E}_{1st}(n_s) = n_I \bar{N} = n.$$

Para obtener una comparación justa, consideremos el muestreo aleatorio simple sin reemplazamiento directo, con el tamaño muestral (fijo)  $n = n_I \bar{N}$ . El estimador de Horvitz-Thompson de  $Y_U$  es entonces  $N \bar{y}_s$ , y la varianza es

$$V_{srswor} = \mathbb{V}_{srswor}(N \bar{y}_s) = \bar{N} K_I S_{yU}^2.$$

Por tanto, una tercera expresión para  $V_{1st}$  es

$$V_{1st} = \left( 1 + \frac{N - N_I}{N_I - 1} \delta \right) V_{srswor} + K_I Cov, \quad (7.24)$$

a través de la cual obtenemos el efecto de diseño de 1st:

$$\text{deff}_{1st} = \frac{V_{1st}}{V_{srswor}} = 1 + \frac{N - N_I}{N_I - 1} \delta + \frac{Cov}{\bar{N} S_{yU}^2}. \quad (7.25)$$

Estas expresiones dan lugar a algunas conclusiones interesantes sobre la eficiencia del muestreo 1st.

*Caso 1*

Supongamos que todos los tamaños de los conglomerados  $N_i$  son iguales, es decir,  $N_i = \bar{N}$  para cada  $i$ . En este caso,  $Cov = 0$ , y obtenemos

$$\text{deff}_{1st} = \frac{V_{1st}}{V_{srswor}} = 1 + \frac{N - N_I}{N_I - 1} \delta \approx 1 + (\bar{N} - 1)\delta. \quad (7.26)$$

Esto muestra que  $V_{1st} < V_{srswor}$  si y solo si  $\delta < 0$ , es decir, si y solo si hay una variación suficientemente grande dentro del conglomerado. Sin embargo, muchos conglomerados con los que trabajamos en la práctica están formados por elementos 'próximos', y, como estos elementos tienen a asemejarse entre sí más o menos, es probable que  $\delta > 0$ .

En consecuencia,  $V_{1st}$  es mayor que  $V_{srswor}$ , a menudo considerablemente mayor. Por ejemplo, incluso si  $\delta$  es positivo pero muy cercano a cero, digamos por ejemplo que  $\delta = 0,08$ , tenemos, con un tamaño medio de conglomerado de  $\bar{N} = 300$ ,

$$\text{deff}(1st, \hat{Y}_U^{\text{HT}}) \approx 25.$$

Esto muestra una gran pérdida de eficiencia debida al muestreo por conglomerados, porque el tamaño medio del conglomerado es bastante grande en este caso.

*Caso 2*

Supongamos que los conglomerados varían en tamaño. Si la correlación entre  $N_i$  y  $A_i = N_i \bar{y}_{U_i}^2$  es positiva, como suele ocurrir a menudo, la varianza aumenta debido a que la selección de conglomerados puede ser peor que en el Caso 1, ya que el segundo término en la fórmula (7.24) puede ser grande.

Para destacar el efecto de la variación en el tamaño de los conglomerados, consideremos el caso extremo de homogeneidad mínima, es decir,  $\delta = \delta_{\min} = -\frac{N_I - 1}{N - N_I}$ . En este caso, todos los  $\bar{y}_{U_i}$  son iguales a  $\bar{y}_U$  y (7.24) se puede escribir como

$$V_{1st} = \bar{y}_U^2 K_I S_{NU_I}^2 \quad (7.27)$$

que será mayor si la varianza del tamaño de conglomerado

$$S_{NU_I}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (N_i - \bar{N})^2$$

es grande.

En este caso

$$\text{deff}_{1st} = \frac{V_{1st}}{V_{srswor}} = \bar{N} \left( \frac{cv_n}{cv_y} \right),$$

donde los dos coeficientes de variación son  $cv_n = \frac{S_{NU_i}}{\bar{N}}$  y  $cv_y = \frac{S_{yU}}{\bar{y}_U}$ . El cociente  $\frac{V_{1st}}{V_{srswor}}$  puede ser mayor que la unidad, especialmente si  $\bar{N}$  es grande.

Esta discusión muestra que la estrategia  $(1st, \hat{Y}_U^{HT})$  es probable que sea ineficiente en muchas situaciones, especialmente si los conglomerados son homogéneos y/o de distintos tamaños. Sin embargo, desde un punto de vista del coste/eficiencia, la estrategia  $(1st, \hat{Y}_U^{HT})$  puede tener ventajas, ya que a menudo es más barato encuestar conglomerados de elementos que encuestar a una muestra dispersa geográficamente que se puede obtener a partir de una selección aleatoria simple de elementos.

Sin embargo, la eficiencia del muestreo por conglomerados se puede mejorar cuando se dispone de información auxiliar. La elección de la estrategia, entonces, depende de la información disponible.

Un caso sencillo surge cuando una medida aproximada  $u_i$  del tamaño  $Y_{U_i}$  está disponible para cada conglomerado  $i = 1, \dots, N_I$ . Si  $u_i$  es aproximadamente proporcional a  $Y_{U_i}$  se puede reducir la varianza del estimador de Horvitz-Thompson usando muestreo por conglomerado *ppswor* con probabilidades de inclusión  $\pi_{Ii} \propto u_i$ . Una alternativa es usar muestreo por conglomerado estratificado con estratos de conglomerados formados de forma que la variación de  $u_i$  sea pequeña en cada estrato.

*Ejemplo 44.* Consideremos dos poblaciones artificiales, cada una con tres conglomerados y tres elementos por conglomerado.

	Población A			Población B		
Conglomerado 1	10	20	30	9	10	11
Conglomerado 2	11	20	32	17	20	20
Conglomerado 3	9	17	31	31	32	30

Los elementos son los mismos para las dos poblaciones, de modo que éstas comparten los valores  $\bar{y}_U = 20$  y  $S_{yU}^2 = 84,5$ . En la población A la mayor parte de la variabilidad aparece dentro de los conglomerados, en la población B la mayor parte de la variabilidad aparece entre los conglomerados.

	Población A		Población B	
	$\bar{y}_{U_i}$	$S_{yU_i}^2$	$\bar{y}_{U_i}$	$S_{yU_i}^2$
Conglomerado 1	20	100	10	1
Conglomerado 2	21	111	19	3
Conglomerado 3	19	124	31	1

Obtenemos la Tabla 7.1 de análisis de la varianza para la población A y la Tabla 7.2 de análisis de la varianza para la población B.

Fuente	gl	SC	CM	F
Entre los conglomerados	2	6	3	0.03
Dentro de los conglomerados	6	670	111.67	
Total	8	676	84.5	

Tabla 7.1: Tabla ANOVA de la población A.

Fuente	gl	SC	CM	F
Entre los conglomerados	2	666	333	199.8
Dentro de los conglomerados	6	10	1.67	
Total	8	676	84.5	

Tabla 7.2: Tabla ANOVA de la población B.

De esta forma tenemos:

$$\delta = -0,3215 \text{ para la población A.}$$

y

$$\delta = 0,9803 \text{ para la población B.}$$

La población A tiene mucha variación entre los elementos que se encuentran dentro de los conglomerados, pero poca variación entre las medias de los conglomerados. Esto se refleja en el valor negativo de  $\delta$ . Los elementos del mismo conglomerado son, en realidad, menos similares que los elementos elegidos al azar de toda la población. Para esta situación, el muestreo por conglomerados es más eficiente que el muestreo aleatorio simple.

Lo contrario ocurre en la población B: la mayor parte de la variabilidad ocurre entre los conglomerados, y éstos son relativamente homogéneos. El valor de  $\delta$  es muy cercano a 1, lo cual indica que se obtiene poca información nueva al incluir en la muestra a más de un elemento del conglomerado. En este caso, el muestreo por conglomerados en una etapa es mucho menos eficiente que el muestreo aleatorio simple. ■

### **Comentario 39. Muestreo por conglomerados sin submuestreo y muestreo por conglomerados con submuestreo.**

Hemos visto que en el muestreo por conglomerados en una etapa se incluyen en la muestra todas las SSUs dentro de la PSU seleccionada. Sin embargo, en muchas situaciones, como por ejemplo la población B del Ejemplo 44, pueden ser demasiado similares,

de modo que el análisis de las SSUs dentro de la PSU será un desperdicio de recursos.

Alternativamente, puede resultar muy caro medir una SSU con respecto al coste de una PSU. En estos casos, podría ser más barato tomar una submuestra de SSUs dentro de cada PSU. Esto se verá ya en el Tema 1 del bloque de 'Producción Estadística Oficial: Métodos Avanzados de la especialidad de Estadística-Ciencia de datos'. ■

## Bibliografía

- Basow, S.A. y N.T. Silberg (1987). "Student evaluations of college professors: Are female and male professors rated differently?" En: *Journal of Educational Psychology* 79, págs. 308-314.
- Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.
- Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

## Tema 8

**Métodos y gestión de la recogida de datos. Introducción. Métodos de recogida de datos básicos. Recogida de datos asistida por ordenador. Otros métodos de recogida. Introducción. Implementación de la encuesta. Gestión activa del trabajo de campo. Paradatos. Monitorización de la calidad de la respuesta. Monitorización del proceso de producción de una encuesta. Evaluación de la encuesta y el informe de calidad.**

Este tema está elaborado usando la siguiente bibliografía.

G. Snijkers y col. (2013). *Designing and Conducting Business Surveys*. Wiley

Eurostat (2017). *Handbook on Methodology of Modern Business Statistics (Memoboost)*.  
URL: [https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics\\_en](https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en)

A. Wallgren y B. Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. 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.

### 8.1 Introducción a los métodos de la recogida de datos

La recogida de datos es un “proceso sistemático de recogida de datos para estadísticas oficiales”, según se recoge en el *Statistical Data and Metadata eXchange* (SDMX 2012)<sup>1</sup>.

Además, el proceso de recogida de datos está muy articulado, y que se desarrolla a lo largo de diferentes fase del proceso de producción: desde la fase de diseño del cuestionario y de la metodología a emplear, hasta la finalización de la información obtenida

---

<sup>1</sup>sistema de intercambio de datos y metadatos estadísticos, que fue creado en 2001 por siete organizaciones que trabajaban en estadísticas a nivel internacional: Banco de Pagos Internacionales (BPI), Banco Central Europeo (BCE), Eurostat, Fondo Monetario Internacional (FMI), Organización para la Cooperación y el Desarrollo Económico (OCDE), División de Estadística de las Naciones Unidas y el Banco Mundial. Estas siete organizaciones trabajan como patrocinadores de SDMX



(UNECE 2019), definiendo el GSBPM (del inglés *Generic Statistical Business Process Model*, véanse los Temas 15 y 16).

Dicho esto, la recogida de datos se basa en reunir la información necesaria de cada unidad seleccionada en la muestras, considerando que existen diferentes fuentes de datos: a) encuestas (la más utilizada tradicionalmente); b) registros administrativos (RR.AA., véase Tema 14); c) datos digitales. Cabe señalar que la recogida de datos es diferente para cada tipo de fuente y que se utilizan diferentes técnicas que pueden, o no, ser asistidas por ordenador y pueden, o no, necesitar el apoyo de entrevistadores (entre las principales podemos señalar: CAPI, CATI, CAWI, PAPI, cuestionarios por correo y observación directa). En este tema nos centraremos en la recogida de datos a través de encuestas.

Durante la recogida de datos, los miembros de la población son localizados, se contacta con ellos y se busca su participación en la encuesta. A continuación se usa un cuestionario y se graban sus respuestas. Este proceso es caro, requiere mucho tiempo y muchos recursos, y tiene un impacto directo en la calidad de los datos. El (Los) método(s) de recogida de datos deben de elegirse para conseguir una tasa de respuesta alta y recoger los datos lo más completos y acurados posible, tratando de minimizar la falta de respuesta.

Hay que aclarar que habitualmente se utiliza el término “encuestado (s)”, el cual, pretende representar a todos los “actores” involucrados en brindar la información a recolectar de acuerdo a las necesidades de las encuestas. Por tanto, los encuestados se pueden definir como “empresas, establecimientos, instituciones, personas individuales, etc., de quienes se recopilan datos e información asociada para su uso en la compilación de estadísticas”, (Glosario de términos estadísticos de la OCDE <sup>2</sup>). Es decir, en esta definición se incluyen todas las expresiones como “unidades informadoras”, “unidades de observación”, “proveedor de datos”, etc. cuya definición se puede encontrar en dicho glosario.

Los paradata juegan un papel muy importante de la recogida de datos. Se definen los paradata como la información relativa a la recogida de datos o al proceso de producción estadístico [...] distinto de la información que es el objetivo de la recogida de datos o al proceso de producción estadístico. Algunos ejemplos de paradata es la duración de la encuesta, el estado de la encuesta (cuestionario enviado, cuestionario grabado, cuestionario depurado), los edits que no se verifican, las observaciones de los informantes, etc. Los paradata se verán en la Sección 8.8.

## 8.2 Métodos de recogida de datos básicos

En relación con el proceso de recogida de datos y con los diferentes métodos existentes, indicar que la elección de la técnica a utilizar depende de muchos factores (tema de la

---

<sup>2</sup><https://stats.oecd.org/glossary/>

encuesta, tiempo disponible para la recogida, dificultad para encontrar la información requerida, tipo de informantes involucrados, presupuesto, etc.) y generalmente se decide durante la fase de diseño del proceso, ya que el modo de recogida influye en la forma en que se lleva a cabo la recogida de datos, así como en el diseño del cuestionario de la encuesta. El uso del modo mixto <sup>3</sup>, es decir, la combinación de diferentes modos de recogida de datos para la misma encuesta, puede superar las limitaciones específicas de cada técnica y, si se diseña correctamente, puede reducir la tasa de falta de respuesta de la operación estadística.

Una tendencia general entre los Institutos de Estadística es recoger la información que necesitan mediante el uso de datos administrativos para reducir la carga a los informantes y el coste. Esto se debe a que los INEs pueden aprovechar los datos ya existentes, almacenados en registros administrativos en poder de otros organismos públicos que ya han realizado una fase de recogida de datos, de acuerdo con sus necesidades y propósitos, aunque en algunos casos, las definiciones administrativas pueden diferir de las estadísticas.

El Intercambio Electrónico de Datos (EDI del inglés *Electronic Data Interchange*) representa otra forma de recoger datos tanto de las instituciones públicas como de las empresas, ya que se basan en el intercambio de información entre el proveedor de datos y el INE sobre la base de un sistema común y modelo de datos estructurados acordado.

El proceso de recogida de datos no es sólo una cuestión de técnicas de entrevista, sino también de estrategias de contacto, así como de actividades de seguimiento. El primer conjunto de actividades es necesario para ponerse en contacto con los informantes y puede variar según el tipo de unidad estadística (empresas nuevas, personas, etc.). El segundo conjunto de actividades es importante para mantener bajo control la recogida de datos mientras tiene lugar y una vez recogidos los datos, para tomar las medidas necesarias para asegurar la calidad de los datos recogidos.

Durante la recogida y después de la misma, tiene lugar el proceso de depuración de los datos, que permite, mediante la detección y corrección de errores, evitar sesgos e inconsistencias. La depuración de los datos se ve en el Tema 9 y se condiciona por el/los modo(s) utilizado(s) durante la recogida.

Los métodos de recogida de datos básicos son:

- Autocumplimentación. En este caso los informantes rellenan el cuestionario (en papel, en formato electrónico u online) sin la ayuda de un entrevistador. El cuestionario se puede enviar y recibir por correo postal o electrónicamente, o estar disponible online. Estos métodos requieren un cuestionario muy bien estructurado y fácil de seguir, con instrucciones muy claras. Se suele proporcionar un teléfono gratuito y un mail de contacto para ayudar al informador. Es más barato que los

---

<sup>3</sup>del inglés *mixed mode*

métodos que usan un entrevistador. Se suele usar para encuestas que precisan información detallada, ya que el informante puede consultar documentos. También es útil para el caso de temas sensibles, ya que el cuestionario puede rellenarse con privacidad. Un inconveniente es que es necesario que los informantes sean expertos en la materia y es necesario disponer de datos de contacto (dirección postal, teléfono, correo electrónico).

- Con la ayuda de un entrevistador. Son la encuesta personal y por teléfono. El mayor beneficio de estos métodos es que un entrevistador puede aumentar la tasa de respuesta (estimulando el interés en la encuesta o tranquilizando a los informantes sobre cualquier duda que pueda tener) y la calidad global de los datos (evitando la falta de respuesta total o parcial, proporcionando las definiciones y conceptos correctos). Estos métodos son útiles cuando los conceptos de los cuestionarios son complejos o cuando la autocumplimentación es difícil. Otra ventaja es que permite periodos de recogida más flexibles. El principal inconveniente es que son métodos caros (especialmente en el caso de entrevistas personales y/o aquellas en las que la muestra esté muy dispersa) y si los entrevistadores no están bien entrenados pueden dar lugar a errores de respuesta. Las entrevistas personales se suelen usar para encuestas muy complejas o cuestionarios muy largos. Las entrevistas por teléfono son más baratas que las personales y es el método más rápido, el principal problema es que es necesario tener un teléfono, por eso se suelen usar después de la personal o de la autocumplimentación.

La siguiente tabla compara los métodos de recogida de datos: entrevistas personales, por teléfono y Auto-cumplimentación:

Comparación de métodos			
	Autocumplimentación	Con entrevistador	
		<b>Teléfono</b>	<b>Personal</b>
Coste	Bajo	Medio	Alto
Tiempo	Más largo	Más corto	En la media
Tasa de respuesta	Bajo	Medio-alto	Alto

### 8.3 Recogida de datos asistida por ordenador

Uno de los principales factores a tener en cuenta en el diseño de la encuesta es si la recogida implica métodos con ordenador o un enfoque tradicional basado en la recogida en papel. Los primeros son cada vez más usados, tanto en las encuestas económicas como en las sociales.

Para encuestas que se van a realizar una única vez los métodos basados en el papel son a menudo más baratos, aunque una vez recogidos es necesario grabarlos y transformar las respuestas en un formato legible por la máquina (con la posibilidad de cometer errores de grabación). El principal beneficio de los métodos asistidos por ordenador es que la recogida y la grabación de datos tienen lugar a la vez, dando lugar a un proceso

de recogida integrado más rápido y eficiente.

Este tipo de métodos se puede llevar a cabo con métodos autocumplimentación (CASI), por teléfono (CATI), en persona (CAPI) u online (CAWI).

El CAWI es un método muy utilizado en el cual el cuestionario, junto con un conjunto de edits que identifican valores missing, inválidos o inconsistentes, está disponible en la página web del instituto de estadística. También podría ocurrir que el cuestionario en formato electrónico se envíe al informante para que lo complemente en su ordenador (CASI). Las ventajas del uso de Internet es que se reducen los costes y se mejora el timeliness (tiempo que pasa entre que ocurre el fenómeno que queremos estudiar y la publicación de los resultados). Entre las ventajas se encuentran la flexibilidad y la comodidad para el informante, ya que puede contestar el cuestionario en varias veces, grabando las respuestas cada vez y retomando el cuestionario donde se había quedado la última vez. Normalmente los informantes reciben un código para poder entrar en el cuestionario y rellenarlo. Hace años el acceso a Internet era un problema que con el tiempo ha desaparecido.

En el caso de CATI y CAPI cada entrevistador tiene un ordenador. El entrevistador lee las preguntas e introduce las respuestas directamente en el ordenador.

El proceso se explica de forma detallada a continuación. Para poder realizar entrevistas CATI, los entrevistadores tienen asignado un grupo de informantes, a los que llamarán y entrevistarán por teléfono, grabando sus respuestas en el sistema de gestión de la recogida de datos. Para poder hacer esto, se deben seguir varios pasos. Uno importante es que los entrevistadores deben de tener un cupo dentro de su horario, los datos de contacto del informante (empresa, establecimiento, individuo) y la hora del día en la que conviene hacerlo preferentemente para aumentar la tasa de respuesta. En ocasiones puede que la información de contacto del encuestado se haya quedado desactualizada, por lo que es conveniente que en el cuestionario haya un apartado para que se proporcionen los datos correctos. Y en el sistema de gestión de la recogida se debe de disponer de un apartado para que los encuestadores incluyan información sobre el mejor horario y forma de contacto (mail, teléfono) con el informante. Es necesario hacer una asignación de encuestadores a encuestados y encuestas. Para ello hay que tener en cuenta los plazos de realización de las encuestas, que varían dependiendo de si las operaciones son coyunturales o estructurales.

Cabe señalar que se pueden combinar los métodos de recogida y proporcionar al informante la posibilidad de usar el que considere en cada momento. Sobre todo en el caso de las operaciones coyunturales o encuestas repetidas en las que el informante se mantiene en la muestra durante largos periodos de tiempo.

En general los métodos asistidos por ordenador tiene muchas ventajas sobre los métodos basados en el papel. Los primeros incluyen edits que permiten la depuración de los

datos en el momento de la recogida, lo que reduce la carga al informante y el tiempo de recogida. Y es más sencillo implementar el control de calidad debido a los parámetros que se generan durante la recogida. Además el cuestionario puede ser más complejo en términos de edits y de saltos de preguntas.

Uno de los inconvenientes de estos métodos es que es necesario la programación de los cuestionarios y de los edits, de la ayuda online, definiciones y descarga de las respuestas en una base de datos. Sin embargo, para operaciones con una muestra muy grande o que sean periódicas, a largo plazo puede resultar más barato.

## 8.4 Otros métodos de recogida

### 8.4.1 Observación directa

Este método consiste en la observación o medición directa de ciertas características de interés. Puede ser la única posibilidad para algunos conceptos (p.e., algunos datos médicos) y es muy común en el caso de encuestas de precios desde la óptica de la demanda. La observación directa se caracteriza por ser no intrusiva, es decir, el elemento observado no interactúa con ningún observador. Eso significa que los datos obtenidos a través de este método son reconocidos y valorados, por su confiabilidad.

Como ejemplos prácticos, los sensores remotos son una forma de observación directa que interpreta imágenes de satélite (p.e., en el censo agrario para estimar superficies); los equipos aforadores (conteo) instalados en centros comerciales, estaciones de autobús, aeropuertos, instalaciones deportivas o de ocio,... para cuantificar el número de personas que entran o salen en determinados períodos temporales (por hora, por día, por semana,...), pueden ser por infrarrojos o mediante cámaras inteligentes.

No se puede aplicar a la mayoría de los datos, ya que no son directamente observables, pero sí a los datos básicos que son necesarios para cuantificar y analizar variables muy concretas (precios, elementos, cantidades). Cuando hay una observación directa no hay falta de respuesta, pero en el caso de la medición, la carga al informante puede ser alta. Uno de los inconvenientes es que puede resultar cara debido al tipo de equipo de medición necesario.

### 8.4.2 Intercambio electrónico de datos (EDI, *Electronic Data Interchange*)

Algunas encuestas permiten que los datos se descarguen directamente de una base de datos y se proporcionen en un fichero XML (eXtended Markup Language) o XBRL (eXtended Business Reporting Language) que son estándares globales para el intercambio de información financiera o estadística. Se utilizan, normalmente en encuestas económicas. Los lenguajes XML / XBRL etiquetan la información individual en un documento con la información necesaria. Por tanto, es necesaria una taxonomía XML / XBRL (desarrollado por los INEs) que defina las variables y las relaciones que pueden existir entre esas variables. Una taxonomía, a su vez, es el diccionario o el vocabulario

común y sus reglas de uso en XBRL. Específicamente son esquemas o mecanismos de clasificación que definen nombres o etiquetas específicas para cada elemento de información (por ejemplo: “Utilidad Neta”).

El fichero es transferido del ordenador del informante al sistema de gestión de la oficina de estadística mediante un web service o subiendo el fichero a una página web. Entre sus ventajas están que la carga al informante es casi nula, los datos no contienen errores (en caso contrario no se permite su envío) y se reduce el tiempo de respuesta y de procesamiento, además de permitir el envío de una cantidad de datos mayor que con un cuestionario. La principal desventaja es su coste tanto para el informante como para la oficina de estadística.

Por tanto, EDI representa el “intercambio electrónico de datos, generalmente en formas que son compatibles, de modo que el software o una combinación de personas y software pueden poner los datos en una forma compatible en el extremo receptor si es necesario”. (SDMX, 2009). De esta manera, EDI ofrece a las empresas la oportunidad de recuperar información electrónicamente desde sus sistemas internos y para reenviar esa información a socios comerciales/proveedores/clientes/gobierno a través de una red de comunicaciones (SDMX, 2009).

#### VENTAJAS:

- Para los encuestados:
  - Una gran reducción de la carga de respuesta;
  - Una interacción mejorada con los INE.
- Para los INE:
  - Un aumento de la calidad de los datos (sin errores tipográficos, sin falta de datos, validación antes de incluir los datos en la base de datos del INE);
  - Una reducción en los costes (no es necesario grabar los datos, no es necesario depuración o imputación);
  - Una mejora de los tiempos fijados y la puntualidad.

#### DESVENTAJAS:

- Para los encuestados:
  - Coste en dinero y/o tiempo para comprar/developar la aplicación;
  - Los establecimientos pueden tener miedo a los problemas técnicos o de confidencialidad.
- Para los INE:

- Costes de implementar el nuevo sistema de recogida, aunque una vez el sistema está implementado se puede utilizar para otras encuestas, y el uso del software de fuentes abiertas puede ayudar.

### 8.4.3 Datos administrativos

Los datos administrativos son “el conjunto de unidades y datos derivados de una fuente administrativa”, y una fuente administrativa es “un depósito de datos que contiene información recopilada y mantenida con el propósito de implementar una o más regulaciones administrativas” (SDMX, 2009).

Un registro administrativo almacena registros de todos objetos a administrar, siendo requerido que sea posible identificar todos los objetos. Las identidades utilizadas pueden ser números de identidad únicos, dentro de un sistema administrativo nacional, o un número de identidad en un subsistema con claves de las identidades en otros sistemas. Estas identidades se utilizarán en la coincidencia exacta de los objetos en diferentes registros. Un registro estadístico se basa en datos de registros administrativos que se hayan tramitado para adaptarse a fines estadísticos. Dichos registros pueden basarse en un registro de población realizado por el INE, o en registros administrados provenientes autoridades y organizaciones fuera del INE.

Algunas encuestas pueden obtener la información necesaria de datos administrativos existentes. Los datos administrativos son aquellos que han sido recogidos por motivos administrativos (p.e., para administrar, regular o gravar actividades de empresas o individuos) en contra de los motivos estadísticos. Los registros administrativos tienen la gran ventaja de que eliminan una gran parte de los costes de recogida de datos y la carga la informante. También pueden reducir el tiempo de proceso, si los datos necesarios ya existen. Sin embargo, los conceptos y las definiciones deben de ser cuidadosamente evaluados ya que pueden diferir de los estadísticos. También hay una falta de control sobre la calidad de los datos y un trabajo de procesamiento de los datos para que tengan el formato necesario.

Algunos ejemplos de datos administrativos se enumeran a continuación, en base a su tipología:

- (a) Por propietario:
  - (a1) Estado: Registro de población, Registro fiscal, Registro de la propiedad, Registro de la Seguridad Social, empresas de transporte de pasajeros, etc.;
  - (a2) Comunidades Autónomas (con poderes transferidos): Registro Fiscal en País Vasco y Navarra, establecimientos de alojamiento colectivo, empresas de transporte de pasajeros, hospitales, centros educativos, etc.
- (b) Por propósito:

- (b1) Encuestas sociales: padrón de población, catastro, Registro de la Seguridad Social, etc.;
- (b2) Encuestas económicas: Registro Fiscal, Registro de la Seguridad Social, establecimientos de alojamiento colectivo, concesiones de empresas de transporte de viajeros, etc.

En cuanto al uso de los datos Administrativos, los INE pueden usarlos para las siguientes aplicaciones estadísticas:

- en la producción de estadísticas como sustitutos de los datos de las encuestas;
- como input para registros estadísticos que se utilizarán como marco muestral y fuente de información auxiliar en el diseño muestral;
- como fuente de variables adicionales que se utilizarán para estimaciones;
- como información auxiliar para apoyar el procesamiento de datos primarios (por ejemplo, depuración de datos, imputación, calibración de estimaciones).

Los INEs que quieran utilizar datos administrativos deben: (a) Tener conocimiento de la existencia y acceso a los datos administrativos. El derecho de acceso debe estar respaldado por ley así como la protección de la privacidad; (b) Evaluar la idoneidad de su uso (por ejemplo, calidad, puntualidad, definiciones) de los datos administrativos disponibles. Es necesaria la coherencia en cuanto a poblaciones y variables. Una vez que se ha evaluado la aptitud para el uso, es necesario establecer acuerdos de entrega de los datos con el proveedor de los mismos.

#### 8.4.4 Modos combinados (Mixed-modes)

A menudo la estrategia de recogida más satisfactoria es la de ofrecer a los informantes una selección de métodos de recogida. Las ventajas son la mejora de las tasas de respuesta, menos errores de respuesta y una recogida más rápida. Uno de los inconvenientes es que la recogida puede ser más compleja y cara y produce datos heterogéneos que pueden complicar el procesamiento y análisis. Y se puede introducir sesgo si los distintos métodos tienen distinta calidad. Para más información sobre los posible errores en caso de modos combinados, véase la Sección [14.2](#).

En la actualidad muchos INE de todo el mundo están pasando del diseño unimodal al modo mixto. Cuando hablamos de modos en las encuestas, podemos distinguir entre los modos (o canales) de comunicación que se utilizan en una estrategia de comunicación de contacto o encuesta (como cartas de recordatorio y anticipado, seguimiento telefónico, etc.) y modos de recogida de datos, es decir, el modo utilizado para el cuestionario o la entrega de los datos.

Por lo que hay que partir de los diseños de modo mixto y unimodal, discutiendo las características de los distintos modos de recolección de datos.



- Cuestionarios en papel (enviados y devueltos por correo o por fax), los modos tradicionales. Autocumplimentados normalmente. Riesgos de falta de respuesta y errores de medición, alta. Costes, bajos.
- CAWI (Computer Assisted Web Interviewing o encuesta web asistidas por ordenador) son cuestionarios digitales distribuidos a través de medios online que permiten recabar información de manera rápida y analizarla de forma óptima. Se diferencia de las Encuestas CATI y CAPI por ser un proceso menos invasivo e impersonal (especialmente en encuestas a hogares). Autocumplimentada normalmente.
- CATI (Computer-Assisted Telephone Interviewing o encuesta telefónica asistida por ordenador) es un modo muy común en las encuestas sociales desde la década de 1970; en encuestas económicas de modo mixto, se puede utilizar como modo adicional para aumentar las tasas de respuesta o para recontactar durante la fase de depuración. Riesgos de falta de respuesta y errores de medición, baja (dependiendo del tema, puede ser alta en temas sensibles). Costes, medios.
- CAPI (Computer-Assisted Personal Interviewing ó encuesta personal asistida por ordenador), un encuestador (agente de campo) visita a un encuestado para completar o ayudar a completar un cuestionario digitalizado. Riesgos de falta de respuesta y errores de medición, baja (dependiendo del tema, puede ser alta en temas sensibles). Costes, altos.
- TDE (Touchtone Data Entry o entrada de datos por tonos telefónicos), como su nombre indica, las respuestas de la encuesta se contestan utilizando los tonos telefónicos, no siendo necesaria ninguna manipulación de datos al ser recibidos en formato electrónico, sin necesidad de encuestador. Autocumplimentada. Casi no se utiliza en la estadística oficial.
- Smart phone, con cuestionarios web. Autocumplimentada. Es un nuevo método de recogida que se está probando. Plantea el problema de cómo conseguir incluir en la pantalla de un teléfono móvil la pregunta del cuestionario junto con la información auxiliar.

Estos distintos métodos de recogida también permiten clasificar los datos en primarios y secundarios, siendo así considerados en muchos países a la hora de diferenciar los procesos de recogida de datos.

Hay que partir de la idea de que los INEs tienen como objetivo producir estadísticas de calidad y actualizadas, lo que requiere a su vez, datos actualizados en los que se puedan confiar, de fuentes oficiales o contrastadas. Estos podrían ser datos que la propia organización recopila (datos primarios) o datos que están disponibles en el mundo exterior (datos secundarios). Estos últimos pueden ser, por ejemplo, fuentes administrativas mantenidas por otras organizaciones gubernamentales, y fuentes identificadas hoy en día como “Big data”, como datos disponibles en Internet y datos generados por sensores, lo cual enlaza a su vez con el proceso de sensorización que se está produciendo en los últimos años. Conscientes del coste y la carga de respuesta que implica la recogida de datos primarios, cada vez más INEs tienen como objetivo maximizar el uso de datos secundarios para la producción de estadísticas. El proceso completo de recogida de

datos ya existentes se denomina generalmente recogida de datos secundarios.

Para poder utilizar datos de fuentes secundarias, los INE deben saber qué fuentes secundarias existen con respecto a su país y si se les permite acceder a ellas de forma regular. A continuación, es necesario determinar la idoneidad para el uso de la fuente de datos para las estadísticas oficiales. Hay muchas formas de determinar esto y los enfoques más importantes se centran en la calidad de los metadatos de la fuente, en la calidad de los datos de entrada y en la calidad de los datos de las estadísticas producidas, así como en la existencia de una metodología para el uso de los datos. Cuando se considera que una fuente de datos secundaria es adecuada para su uso, es necesario establecer acuerdos de entrega con el proveedor de datos. Se considera una buena práctica asignar a un empleado de INE como persona de contacto para la fuente y el proveedor de datos. Para las estadísticas importantes que dependen de la disponibilidad de los datos secundarios, es necesario configurar formas de lidiar con cualquier interrupción o retraso en la entrega. Estos supuestos escenarios alternativos pueden variar desde acciones muy simples, como contactar directamente con el proveedor de datos, hasta el uso de modelos complejos que son capaces de hacer frente a los datos que faltan (falta de respuesta).

Aparte de los datos administrativos, algunos trabajos más recientes también se centran en el uso de fuentes secundarias innovadoras, los denominados Big Data, para las estadísticas, aunque muchos de estos proyectos todavía están en marcha y estas fuentes todavía no se utilizan, masivamente, para las estadísticas oficiales de los INE.

## 8.5 Introducción a la gestión de la recogida de datos

Lo primero que necesitamos son los elementos principales a tener en cuenta en el diseño de los componentes de la recogida de datos: la muestra, el cuestionario (tanto su diseño como su implementación en los distintos modos), el personal formado y la estrategia de comunicación. Una vez que todos los componentes de la encuesta están listos para su implementación y se lanza la encuesta, hay que llevar a cabo una eficiente gestión del trabajo de campo o de la recogida de datos. Esto implica un control y seguimiento del proceso de recogida de datos, de cara a garantizar la calidad de los resultados.

Entonces, es cuando se comprobará si el diseño de la encuesta planificada es eficaz, en particular si los desafíos de calidad discutidos en fases previas se han abordado de manera adecuada. Por tanto, el enfoque de la gestión del trabajo de campo debe ser activo, lo que puede implicar cambiar el diseño preparado durante el trabajo de campo.

Al discutir la planificación de una encuesta se deben seguir los principios de la gestión de proyectos, al asimilar la ejecución de una encuesta a la ejecución de un proyecto. Para ello, se debe planificar una encuesta con anticipación y trabajar en consecuencia, para asegurarnos de alcanzar los resultados de la encuesta dentro del plazo y el presupuesto, siguiendo los niveles de calidad previamente acordados.

Ello supone la gestión de riesgos, en cuanto a su tratamiento y planificación de acciones de mitigación. Además, hay que tener en cuenta que una parte importante del proceso de producción está más allá del control de la organización de la encuesta: el diseño del proceso de respuesta de los elementos encuestados. Porque, aún con una preparación cuidadosa, la efectividad del diseño para obtener respuestas de calidad es incierta. En la etapa de diseño, las pruebas preliminares y la prueba piloto pueden ayudar a estimar estas incertidumbres y mejorar el diseño inicial (véase el tema 11 <sup>4</sup> del bloque “Producción Estadística Oficial: Métodos Avanzados” de la especialidad de Estadística y Ciencia de datos del grupo B de materias específicas). Estas incertidumbres amenazan la capacidad organizativa de las encuestas para cumplir con los resultados marcados inicialmente, considerando el plazo y el presupuesto (Groves y Heeringa 2006; Wagner 2010; Kreuter, M. Couper y Lyberg 2010).

Para que los datos recogidos tengan una buena calidad es necesario controlar el trabajo de campo de forma continua en tiempo real. Esto depende de la disponibilidad en tiempo real de parámetros. Los indicadores basados en parámetros permiten la identificación de problemas en el proceso de recogida de datos y la intervención adecuada para aumentar la probabilidad de lograr los objetivos de la encuesta (Groves y Heeringa 2006; Kreuter, M. Couper y Lyberg 2010). Sin embargo, esto también debe planificarse cuidadosamente.

Por tanto, es necesario realizar una gestión activa de la recogida de datos, partiendo de unos diseños receptivos a los elementos de respuesta, para posteriormente discutir sobre el mejor proceso de implementación de la recogida de datos. Los parámetros y los indicadores de control y seguimiento, lógicamente, están relacionados directamente con los errores en las encuestas, lo cual nos lleva a saber si disponemos de unos datos de calidad. Esto nos proporciona un marco para seleccionar indicadores adecuados para monitorizar la recogida de datos, los cuales pueden estar relacionados con (1) el resultado del trabajo de campo (es decir, los datos de la encuesta) y (2) el proceso de trabajo de campo, que se relaciona con la calidad y el coste del diseño de la encuesta (Renssen y col. 1998):

- indicadores de calidad de los datos de la encuesta, incluida la tasa de respuesta;
- indicadores de la eficacia y eficiencia del diseño de la encuesta (con respecto al coste y los recursos).

## 8.6 Implementación de la encuesta

Antes de que pasar de la etapa de diseño, construcción y prueba (design–build–test, DBT) de la recogida de datos, o implementación de la encuesta, es necesario disponer de

---

<sup>4</sup>Tema 11. Métodos para el desarrollo, testeo y evaluación de instrumentos de recogida de datos I. Un marco para el desarrollo, testeo y evaluación. Desarrollo de contenido, medidas y cuestiones en encuestas. Testeo de preguntas y cuestionarios. Evaluación de preguntas y cuestionarios. Métodos para el desarrollo, testeo y evaluación de instrumentos de recogida de datos. II. Desarrollo, testeo y evaluación de instrumentos de recogida electrónica de datos. Análisis de datos cuantitativos. Enfoques multimétodo para el desarrollo, testeo y evaluación. Organización y logística.

los elementos fundamentales listos: la muestra, el cuestionario, el material de comunicación. Igualmente, se han de tener todos los sistemas implementados, las herramientas desarrolladas, y dispuestos los flujos de trabajo y operacionales relacionados con la recogida de datos, y habrán sido diseñados, contruidos y probados los métodos y procedimientos asociados: grabación, codificación y depuración de datos.

La implementación de la encuesta implica tener todo listo para el trabajo de campo, lo que afecta a todos los componentes del proceso de recogida de datos. Esto incluye:

- La muestra ha sido seleccionada y se dispone de una lista de unidades con información actualizada sobre ellas.
- El cuestionario final está impreso, los cuestionarios electrónicos están disponibles en la página web, los cuestionarios CATI y CAPI están listos en los ordenadores de los encuestadores, y los cuestionarios para otros modos de recogida de datos están implementados en los equipos y el resto de actividades necesarias (el esquema XML/XBRL está disponible para los usuario, se he desarrollado una aplicación para subir los ficheros XML/XBRL, etc.) .
- Se tiene el calendario para la estrategia de comunicación y recogida de datos, todos los documentos de comunicación están impresos, subidos en Internet o implementados, los entrevistadores tienen su listado de informantes.
- Los métodos y procedimientos de grabación, codificación y depuración de datos están implementados en las herramientas e instrucciones correspondientes.
- Todos los sistemas de producción, herramientas y procedimientos están finalizados, incluyendo los sistemas de datos, metadatos y parados.
- El personal ha sido instruido en relación con sus funciones y está capacitado en habilidades, procedimientos, sistemas y herramientas.
- El procedimiento de recogida de datos está montado, al igual que la logística, los flujos de trabajo y las operaciones. La secuencia de actividades destinadas a realizar con éxito la encuesta y cumplir con los objetivos marcados está lista.

Todo esto requiere una planificación previa. Durante la etapa DBT, la planificación de la encuesta puede haber cambiado. En ocasiones puede que el trabajo de campo se retrase, quizás debido al tiempo adicional necesario para diseñar y probar el cuestionario, pero esto no significa necesariamente que la finalización del trabajo de campo se encuentre en una etapa posterior. Y por supuesto, no tiene por qué afectar al resultado final de la encuesta. Sin embargo, si un inicio posterior del trabajo de campo afecta los componentes de la encuesta, estos deben actualizarse. Es posible que sea necesario actualizar la información de contacto, es posible que se necesite establecer un nuevo calendario de trabajos y actividades, lo que afectará al personal involucrado en el trabajo de campo.

Para que el trabajo de campo resulte eficaz, todas las componentes deben de funcionar adecuadamente y de forma sincronizada. En la etapa de diseño, se han probado los componentes, procedimientos, sistemas y herramientas individuales, pero aún no se

han probado como un sistema conjunto en condiciones reales.

Por tanto, es muy recomendable realizar un piloto antes de realizar la encuesta, tomando de ejemplo a una parte de la muestra. Sobre todo, es muy recomendable realizarla, al menos, en caso de un rediseño importante de la encuesta como puede ser el rediseño completo del cuestionario, o un sistema de recogida de datos completamente nuevo. En una encuesta recurrente, esto se puede hacer, por ejemplo, con una parte representativa de la muestra. De esta forma, en lugar de correr el riesgo de que surja un problema con toda la encuesta, el riesgo se reduce a una pequeña parte de la muestra, lo que garantiza la entrega de resultados. Una vez más, la realización de una prueba piloto también requiere una planificación previa (para más información sobre pruebas piloto véase el tema 11 de la especialidad de Estadística y Ciencia de datos del grupo B de materias específicas).

## 8.7 Gestión activa del trabajo de campo

Finalmente, las unidades muestrales son contactadas, la recogida de datos está en funcionamiento y, por tanto, es necesaria una gestión del trabajo de campo ya que nos podemos encontrar con incertidumbres que pueden afectar a los resultados finales. Es imprescindible un sistema eficaz de control y seguimiento de los trabajos que nos permita monitorizar el proceso de recogida de datos, respondiendo a preguntas clave: ¿cuál es la tasa de respuesta? ¿cuál es el edit que mayor número de unidades no verifican? ¿cuál es el modo de recogida más utilizado por las unidades?, ¿es necesario recontactar con las unidades?, ¿cuántas unidades son sancionadas por no proporcionar la información?.

Pero, la gestión no es una única acción, sino una serie de pasos, y para que funcione correctamente la gestión del trabajo de campo debemos adoptar un enfoque activo de dicha gestión. Esto incluye:

- Planificación. Incluye dos aspectos: (1) identificar indicadores relevantes para hacer un seguimiento del proceso en general; y (2) además, identificar características críticas del diseño y planificar intervenciones atenuantes.
- Seguimiento. El objetivo es identificar problemas durante la recogida tan pronto como sea posible.
- Identificación de los problemas y encontrar soluciones. Por ejemplo, si la tasa de respuesta es muy baja, se puede recontactar con las unidades para ver si tienen algún problema en la cumplimentación, si no entienden bien el cuestionario, etc.
- Propuestas de acciones correctivas. Una vez que se detectan los problemas, las propuestas de acciones correctivas (preparadas previamente) son propuestas a los responsables, que tomarán las medidas necesarias.
- Comunicación y toma de decisiones. Una vez tomadas las decisiones es necesario su comunicación a todo el personal involucrado.

- Evaluación y documentación. Una vez el proceso ha terminado tiene lugar la evaluación y la documentación, con el fin de mejorar si es necesario, más adelante.

Es decir, en la práctica, al realizar encuestas hay que solventar muchas incertidumbres que afectan los costes y errores de las mismas, incertidumbres que pueden no haber sido identificadas previamente durante las pruebas preliminares. En la gestión activa del trabajo de campo, el objetivo general del seguimiento es identificar los problemas de recogida durante la recogida de datos lo antes posible en el proceso y no al final, cuando es demasiado tarde para actuar. En este punto la pregunta relevante es, ¿vamos por buen camino?

Para responder a esta pregunta, es importante monitorizar muy de cerca las características críticas del diseño de la encuesta, lo que se denomina *enfoque de diseño receptivo* y fue propuesto por [Groves y Heeringa 2006](#). En este enfoque, el seguimiento está dirigido a las características de diseño potencialmente problemáticas y que previamente identificadas, lo que reduce el riesgo de no cumplir con los objetivos de la encuesta. Este enfoque incluye la planificación de intervenciones apropiadas en el diseño y la planificación de cuatro pasos:

1. Identificación previa de las características del diseño que potencialmente afectan a los errores y los costes de la encuesta, p. ej. basado en pruebas piloto, ya que es probable que estas características de diseño pongan en peligro obtener los resultados previstos.
2. Identificación de indicadores, que muestren las propiedades de coste y error de esas características de diseño, y monitorización de esos indicadores durante la recolección de datos. Para calcular estos indicadores, necesitamos definir los parámetros apropiados.
3. Preparar la recogida y el análisis en tiempo real de los parámetros. Esto implica que durante la recogida de datos, los indicadores relevantes deben calcularse en tiempo real.
4. En las fases posteriores de la recogida de datos, las características de la encuesta se pueden modificar de acuerdo con las “reglas de decisión de compensación de costes y errores”. Las reglas de decisión deben definirse y concretarse antes del trabajo de campo y deben garantizar que se alcancen los niveles de calidad predeterminados de los resultados de la encuesta. Sobre la base de estas reglas de decisión, se deben preparar diseños alternativos y acciones apropiadas.

En los *diseños receptivos*, el diseño de la encuesta se cambia, si es necesario, durante la recogida de datos. El desafío es planificar la recogida de datos de tal manera que el diseño realmente se pueda cambiar. Por tanto, en los *diseños receptivos*, el proceso de recogida de datos se divide en fases, y cada fase se supervisa calculando el conjunto de indicadores definidos inicialmente, de tal manera que los resultados se corresponden con determinadas reglas de decisión. Al final de cada fase, si los indicadores muestran niveles inaceptables de acuerdo con las reglas de decisión, se toman las acciones apropiadas.

En determinadas encuestas, esto incluye, por ejemplo, un cambio en el modo de recogida de datos en subpoblaciones específicas (de la web al papel); en una fase posterior, para el control y seguimiento de la falta de respuesta, se puede implementar un cambio de recordatorios en papel, a recordatorios telefónicos; y en una fase final, el cuestionario puede acortarse con solo preguntas clave básicas ([Bethlehem 2009](#)).

Si se dispone de un cuestionario acotado en número de preguntas clave, se necesitará menos tiempo para la recogida de datos y se podrá realizar más encuestas en menos tiempo, optimizando así los recursos. Esto requiere una planificación anticipada y un proceso de recogida de datos flexible.

Se pueden encontrar ejemplos de reglas de decisión en los Estándares de Calidad de las oficinas de estadística o en organismo internacionales (como Eurostat). A modo de ejemplo, el subrequisito establecido en EE.UU. indica que “los análisis de sesgo por falta de respuesta deben realizarse cuando las tasas de respuesta por unidad, artículo o cantidad (es decir, ponderadas) para la muestra total o subpoblaciones importantes caen por debajo de los siguientes umbrales”: (1) El umbral para las tasas de respuesta unitaria es del 80 %; (2) El umbral para las tasas de respuesta de los elementos clave es del 70 %; (3) El umbral para las tasas de respuesta de cantidad total es del 70 %.

Teniendo en cuenta que los umbrales 1 y 2 no se aplican a las encuestas que utilizan tasas de respuesta de cantidad total.

## 8.8 Paradatos

Los parados, o datos de procesamiento, son esenciales para la gestión del proceso de recogida de datos. Con el enfoque de la gestión activa del trabajo de campo, permiten el seguimiento y la toma de decisiones para este proceso. El concepto de parados fue introducido por [M.P. Couper 1998](#). Los parados incluyen una gran variedad de datos. Un gran número de clasificaciones de parados han sido propuestos, por ejemplo, dependiendo del tipo de datos o macro/microparados. Entre los objetivos de los usos de los parados están los siguientes:

- Seguimiento y evaluación de una entrevista individual. Estado de la recogida de los datos, de los edits, información acerca del informante.
- Seguimiento y evaluación del marco de producción de la encuesta. Mejoras a corto plazo (feedback continuo con el proceso de producción), mejoras a largo plazo (para evaluar la eficiencia y calidad a lo largo del tiempo para mejorar el proceso), mejoras metodológicas (investigación con el fin de mejorar las componentes de la encuesta como el cuestionario o los edits).
- Imagen corporativa. Proporcionando al público general indicadores de calidad, para alcanzar la confianza en los resultados estadísticos.

El uso de parados en el contexto de la gestión activa del trabajo de campo para el

seguimiento de encuestas individuales requiere un enfoque conjunto en el proceso y la gestión de la calidad. Desde la perspectiva de la gestión del proceso, los parados están relacionados con varias tareas en la recogida de datos y sus costes:

- Parados asociados con el proceso en la organización. Logística (cuándo se envió el cuestionario), actividad del personal (cuándo y cómo contactar con el informante), proceso de grabación de los datos (número de cuestionarios procesados), estado de la respuesta (tasa de respuesta, tiempo de respuesta), proceso de depuración (variables y registros que han sido combinados, número de recontactos).
- Parados asociados con el proceso desde el lado del informante. Resultado mientras se contesta el cuestionario (número de cuestiones modificadas, edits que han saltado), resultado del proceso de realización (estado de la respuesta), help desk, información para el acceso al cuestionario (código y contraseña).
- Parados de gestión del proyecto. Gasto de dinero, recursos humanos y de material, y de tiempo.

Por último, los parados están relacionados con las causas de error en las encuestas. Estos parados y sus indicadores se pueden usar para evaluar la calidad de los componentes de la encuesta y de los datos. Algunos ejemplos son:

- Errores de muestreo: errores de cobertura. Error de sobrecobertura: número de duplicados, de erróneamente incluidos, de unidades no localizables.
- Errores de muestreo: errores de falta de respuesta. En la salida a campo: número de unidades no contactadas. Durante la recogida: recordatorios enviados, número de llamadas, sanciones. Depuración: número de recontactos.
- Errores de medida. Modo de recogida: tasa de respuesta para cada modo, cambios de modos. Carga al informante: tiempo necesario para completar el cuestionario, número de informantes. Cuestionario: falta de respuesta parcial, preguntas con mucha falta de respuesta.

## 8.9 Monitorización de la calidad de la respuesta

Durante la recogida de datos necesitamos realizar un seguimiento de la respuesta para asegurarnos de que los niveles de calidad son los deseados (para más información sobre la calidad véase el Tema 18). Los indicadores de calidad del producto final incluyen:

- A nivel de unidad: tasa de respuesta, de falta de respuesta y de devolución de cuestionarios. En el caso de la tasa de devolución se considera en el numerador el número de unidades que enviaron el cuestionario. Estos indicadores se pueden calcular para cada uno de los modos de recogida. Estos indicadores también se pueden calcular para tipos específicos de subpoblaciones.
- En relación con la importancia relativa de la unidad: tasa de respuesta ponderada. En algunas operaciones estadísticas, como pueden ser las encuestas económicas, algunas empresas o establecimientos son más relevantes que otros. Un pequeño número de unidades puede contribuir a la estimación en mayor proporción. Por



eso puede ser más significativa la tasa de respuesta ponderada. En los pesos se puede considerar distintas variables, como la cifra de negocios o el número de empleados de la empresa o del establecimiento.

- En relación con la distribución de la respuesta: indicador de representatividad. Después de monitorizar la tasa de respuesta ponderada, necesitamos monitorizar la representatividad de la distribución de respuesta para las variables auxiliares.
- Sobre la calidad de las medidas: falta de respuesta parcial y el gráfico de eficiencia del proceso. El gráfico de eficiencia del proceso es un gráfico de control que traza la evolución de las estimaciones para una variable objetivo cuando se reciben los cuestionarios. En el momento de empezar la recogida de datos se sitúa en el 100 % de falta de respuesta y posteriormente esta tasa va descendiendo, sobre todo en las fechas en que se mandan los recordatorios y en los que se sanciona a las unidades. Se debería de disponer de la tasa de respuesta ponderada y sin ponderar.

Los indicadores de calidad de los datos indican lo cerca que estamos de la muestra inicial que se extrajo para obtener resultados acurados, completos y a tiempo. Uno de los indicadores de calidad más comunes es la tasa de respuesta (o de falta de) tanto ponderada como sin ponderar, porque es un indicador fácil de calcular y de interpretar. Pero hay que tener en cuenta varios aspectos de la respuesta para tener una idea clara de la calidad de la respuesta.

Si se utiliza una muestra existe una varianza de muestreo (error de muestreo). Esta varianza se controla con el procedimiento de selección de la muestra. La falta de respuesta total introduce una varianza adicional. La falta de respuesta también puede dar lugar a sesgo si la distribución de la falta de respuesta está sesgada para las variables objetivo o auxiliares. El sesgo será más importante cuando más grandes (en peso) sean las unidades que no responden.

Un segundo tipo de falta de respuesta es la parcial. Esto introduce una varianza y un sesgo adicional.

## 8.10 Monitorización del proceso de producción de una encuesta

Además de realizar un seguimiento de la respuesta, será necesario monitorizar también los procesos de campo y posterior con el fin de estudiar la efectividad y el coste-eficiencia del diseño de la encuesta. El diseño de la encuesta es efectivo si produce datos acurados, completos y en plazo. La monitorización del proceso incluye por tanto la monitorización de la muestra, el cuestionario, el modo de recogida de datos, la estrategia de comunicación y el procesamiento de los datos, en relación con las fuentes de error.

Veamos algunos de estos indicadores:

- Muestra.
  - Tasa de ilocalizables.

- Tasa de unidades mal clasificadas.
  - Tasa de duplicados.
- Calidad del cuestionario y de la respuesta.
  - Tasa de recontacto.
  - Tasa de edits que han saltado para las principales variables.
  - Tasa de cambio por pregunta.
- Envío de cuestionarios. Número de cuestionarios enviados.
- Modo de recogida de datos.
  - Tasa de respuesta y de devolución por modo.
  - Tasa de cambio de modo.
- Estrategia de comunicación
  - Tasa de respuesta y de devolución antes del primer recordatorio, o en general, antes de cada recordatorio.
  - Tasa de intento de contacto.
  - Tasa de no contacto.
  - Tasa de negativa a responder.
  - Tasa de aplazamiento.
- Procesamiento de los datos
  - Tasa de no verificación de los edits.
  - Tasa de imputación automática y manual.

## 8.11 Evaluación de la encuesta y el informe de calidad

Después de terminar el trabajo de campo hay que evaluar el proceso estadístico. Esto incluye:

- Una evaluación del proceso de campo. Esta evaluación nos permitirá ver si (1) se han alcanzado los resultados de acuerdo con las especificaciones iniciales, y (2) el trabajo de campo ha sido efectivo a la hora de alcanzar los resultados y dónde se puede mejorar. Esta evaluación puede ser realizada internamente o por auditorías externas.
- Una evaluación del usuario. A los usuarios de los datos se les pregunta por varios aspectos de calidad de los resultados como la relevancia, acuracidad, comparabilidad, oportunidad, etc.
- Una evaluación de la gestión del trabajo de campo. Esto conlleva una evaluación del dinero gastado, los recursos y el tiempo empleados.

- Una evaluación en relación con la carga de respuesta.

Estas evaluaciones se relacionan con los objetivos de mejora del uso de parados. Como consecuencia de estas evaluaciones, se propondrán medidas de mejora: proceso a corto plazo, proceso a largo plazo y/o mejora metodológica. Estas evaluaciones también deben planificarse en el plan del proyecto del trabajo de campo. Los resultados de estas evaluaciones se documentan en un informe de calidad.

Todos los indicadores relevantes definidos inicialmente están documentados en el informe de calidad sobre el proceso de producción de la encuesta, que debería incluir la tasa de respuesta general no ponderada, la tasa de respuesta completa, la tasa de no respuesta y la tasa de retorno. La tasa de respuesta por unidad de tiempo se puede mostrar en gráficos de control y se puede tabular la distribución de las tasas de respuesta para subgrupos o estratos relevantes, al igual que la distribución de respuesta para los diversos modos de recogida de datos. Para mostrar la calidad de los datos, se pueden publicar las tasas de respuesta ponderadas y, para las variables clave, el gráfico de eficiencia del proceso y las tasas de respuesta de los elementos. Es recomendable no solo definir estas tasas, si no también mostrar sus fórmulas descriptivas, incluyendo una tabla que muestre los códigos de resultado de la respuesta final.

Además, el informe debe incluir una sección sobre costes y carga de respuesta, indicando los costes totales de la encuesta, los costes promedio por unidad de respuesta y las cargas de respuesta reales y percibidas, que deberán contrastarse con las estimadas y requeridas para alcanzar los objetivos marcados.

El informe de calidad también debe incluir una sección metodológica que describa el diseño de la encuesta y las medidas que se hayan tomado para asegurar la calidad, incluida la minimización de los errores, actualización del marco muestral que incluya un procedimiento de muestra eficiente, pretest del cuestionario en la etapa de conceptualización y planificación de la operativa, la adaptación del cuestionario, la aplicación del enfoque de diseño de siete pasos para la estrategia de comunicación y la realización de una prueba piloto de campo.

En la página web del INE se pueden consultar, para cada operación estadística, el informe metodológico estandarizado. Por ejemplo, el informe de los Índices de precios industriales: <https://ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=30051>.

## Bibliografía

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. New York: Wiley.  
Couper, M.P. (1998). *Measuring survey quality in a CASIC environment*. URL: [http://www.amstat.org/sections/srms/proceedings/papers/1998\\_006.pdf](http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf).

- Eurostat (2017). *Handbook on Methodology of Modern Business Statistics (Memoboost)*. URL: [https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics\\_en](https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en).
- Groves, R.M. y S.G. Heeringa (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs". En: 169, págs. 439-457.
- Kreuter, F., M. Couper y L. Lyberg (2010). *The use of paradata to monitor and manage survey data collection*. Proceedings of the Survey Research Methods Section, ASA (American Statistical Association). URL: <http://www.asasrms.org/Proceedings/y2010f.html>.
- Renssen, R.H., A. Camstra, C. Huegen y W.J.G. Hacking (1998). *A model for evaluating the quality of business surveys*. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.547.7824&rep=rep1&type=pdf>.
- SDMX (2012). *SDMX 2.1 User Guide*. URL: [https://sdmx.org/wp-content/uploads/SDMX\\_2-1\\_User\\_Guide\\_draft\\_0-1.pdf](https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf).
- Snijkers, G., G. Haraldsen, J. Jones y D.K. Willimack (2013). *Designing and Conducting Business Surveys*. Wiley.
- UNECE (2019). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.
- Wagner, J. (2010). "The fraction of missing information as a tool for monitoring the quality of survey data". En: *Public Opinion Quarterly* 74 (2), págs. 223-243.
- Wallgren, A. y B. Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.

## 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.**

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

T. de Waal, J. Pannekoek y S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Amsterdam: 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 J.G. 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 sobredepuración 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 sobredepuración y la depuración creativa<sup>1</sup> véanse [Granquist 1995](#); [Granquist 1997c](#); [Granquist y J.G. 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 <sup>2</sup>.
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.

## 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.

---

<sup>1</sup>Para 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.

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

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 21****Errores sistemáticos.**

Este tipo de errores puede ocurrir cuando un informante malinterpreta o lee incorrectamente una pregunta. 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.

Un error sistemático bien conocido es el llamado error de unidad de medida. Este error ocurre cuando un informante proporciona el valor de una variable en una unidad de medida errónea. Por ejemplo, supongamos que la cifra de negocios total tiene que ser proporcionada en miles de euros, pero se declara en euros.

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.

Otra forma de detectar los errores sistemáticos es usando edits de razón, que fija unos límites aceptables para un cociente de dos variables. Por ejemplo, considerando el ejemplo anterior del error en la cifra de negocios, si se dispone de la cifra de empleados y esta variable no se ve afectada por el mismo error, el edit basado en el cociente entre la cifra de negocios y los empleados serviría para detectar a unidades que incurran en este error (ya que los valores de los cocientes serían anómalos comparados con los valores de los no erróneos).

Una posible causa de errores sistemáticos que hay que intentar evitar es la causada por un encuestador que ha entendido mal la información a recoger. Por este motivo es muy importante la formación del personal de recogida de datos y que las dudas que se planteen durante la recogida sean transmitidas a los expertos en la materia.

Otros ejemplos de errores sistemáticos son los siguientes: errores asociados con un malentendido sobre los filtros de las preguntas en el cuestionario; errores debidos a la falta de habilidad del informantes para proporcionar información basada en una clasificación/definición estadística específica; errores en signos que pueden ocurrir cuando un informante omite de forma sistemática un signo negativo en alguna variable, por ejemplo los beneficios, que puede ser negativa; valores *missing* sistemáticos como no indicar el valor total para algunas variables.

**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.

Si el error aleatorio no es un valor atípico ni influyente la forma de detectarlo es porque no se verifican alguna regla de depuración. Por ejemplo, si en un cuestionario se indica que la edad es '12' y en estado civil figura 'Casado'.

Hay varios principios básicos en la localización de campos erróneos en un registro con inconsistencias. Los principios más conocidos y más utilizados es el paradigma de [Fellegi y Holt 1976](#).

**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.

El modelo que se utiliza para considerar una observación como *outlier* se refiere a una población base, no al total de la muestra, y a menudo hay distintos modelos

para las distintas subpoblaciones. Por ejemplo un modelo puede ser apropiado sólo para empresas con una actividad económica particular. La división 30 de la CNAE ('Fabricación de otro material de transporte') incluye empresas de construcción naval, aeronáutica y de material ferroviario cuyas cifras de negocios se pueden considerar *outliers* pero que no lo son.

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 [R.J.A. Little y Rubin 2002](#).

#### Definición 22

##### 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  $\xi$  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:

**Definición 23****Ignorable.**

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

**No ignorable.**

Si el mecanismo es NMAR o el parámetro  $\xi$  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.

Cabe señalar que los edits se deben basar en el conocimiento sobre el tema, es decir, sobre el conocimiento de las condiciones sociales y económicas que pueden influir a los informantes y las implicaciones que tienen en la relación entre los apartados del cuestionario. Además, la definición de las regiones de aceptación de los distintos edits debería de estar respaldado por métodos estadísticos específicos. En particular, el análisis de las distribuciones conjuntas puede facilitar en gran medida la especificación de edits adecuados al mostrar relaciones entre las variables (Whitridge y J. Kovar 1990). Los métodos gráficos también pueden resultar útiles.

Veamos a continuación varios ejemplos de clases de edits.

#### Definición 24

##### 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 *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 J.G. 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**

La imputación consiste en asignar un valor a un ítem o un grupo de ítems que previamente no tenía valor o ese valor se consideraba erróneo o no ajustado a la realidad. La

imputación es por lo tanto un proceso por el que se generan valores artificiales y por lo tanto introduce un error de imputación. Sin embargo, este error cuenta con la ventaja de ser medible ya que el especialista puede analizar la precisión de las imputaciones y de esta forma estimar el error de imputación.

La metodología para las estadísticas económicas modernas ([Eurostat 2014](#)) distingue tres grandes enfoques para la imputación. El primero consiste en la imputación deductiva o lógica, que consiste en usar reglas de derivación con la información disponible para estimar el valor *missing*. El segundo consiste en usar reglas de predicción estadísticas para obtener modelos donde calcular imputaciones. El tercer grupo consiste en utilizar unidades similares para imputar con este valores a las unidades donde no hay respuesta. Además de estos tres existe también un enfoque más manual de la imputación que consiste en la imputación por el experto en la materia. Veamos los principales métodos.

#### **Imputación deductiva.**

Este es el método que tiene preferencia sobre todos los demás pero en muchos de los casos no puede ser usado. Se trata de un método especialmente interesante cuando tenemos falta de respuesta parcial ya que utilizando el valor de otros ítems se puede deducir el valor *missing*. Por ejemplo, teniendo la cifra de negocios de España y del extranjero y no teniendo el total podríamos imputar el valor total como la suma de ambos.

#### **Imputación basada en modelos.**

Este enfoque de imputación consiste en encontrar el modelo predictivo adecuado para la obtención de la imputación. El modelo toma por lo tanto la información disponible y puede ser más o menos complejo. Dentro de estas técnicas destacan:

- **Imputación por la media.** Como su nombre indica cada valor *missing* es reemplazado por la media de todos los valores disponibles. Este modelo tiene el problema de no representar la distribución real del fenómeno ya que existirán muchos casos del valor de la media que no se ajustan a la realidad ([Särndal y Lundström 2005](#)). Para reducir esto se pueden realizar imputaciones por media para grupos concretos lo que reduciría este problema.  
Este método tiene la ventaja de ser muy sencillo y no necesitar información auxiliar pero solamente arrojaría resultados satisfactorios a la hora de calcular medias y totales poblacionales pero en ningún caso obtendríamos microdatos ajustados a la realidad. Por otro lado, la existencia de outliers afecta de manera muy negativa a esta técnica ya que las imputaciones se alejaran mas todavía de la distribución real.
- **Imputación por razón.** La imputación por razón tiene en cuenta una sola variable auxiliar y asume que esta variable es proporcional a la variable de estudio. El proceso de estimación consiste en calcular la razón de la variable auxiliar y la variable de estudio sobre el conjunto de datos sin valores *missing*. Una vez calculada esta razón se multiplica por el valor de la variable auxiliar de los valores *missing* y así se obtiene la imputación. Esta estimación será mejor cuanto mayor linealidad

exista entre la variable de estudio y la auxiliar. En este caso al igual que en la imputación por la media se pueden calcular razones para subgrupos dentro de la población si se dispone de más información. El inconveniente es que se necesita información sobre una variable auxiliar en aquellos que han respondido y en los que no lo han hecho. Para la obtención de esta información es habitual hacer uso de los registros estadísticos disponibles o otras operaciones estadísticas sobre la misma muestra.

- **Imputación por regresión.** Esta técnica es una generalización de las dos anteriores para un conjunto de variables auxiliares  $x_1, \dots, x_n$ . El modelo mas simple es el de la regresión lineal que tiene la siguiente expresión:

$$y = \alpha + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon$$

donde  $\alpha$  es el parámetro de la constante y  $\beta_1, \dots, \beta_n$  serán los parámetros desconocidos de cada una de las variables auxiliares,  $\epsilon$  será el error e  $y$  es la variable de estudio. Normalmente la estimación de estos parámetros se realiza mediante mínimos cuadrados ordinarios.

El valor estimado se puede obtener tanto sumando el error como no haciéndolo. Sumar el error no es necesario cuando el objetivo de la imputación es obtener medias o totales poblacionales pero si se quiere observar la variabilidad es conveniente añadir el error (véase [Waal, Pannekoek y Scholtus 2011](#), sección 7.3). El modelo sin error siempre tendrá el mismo resultado y por lo tanto es determinista pero la regresión añadiendo el componente de error será determinista si la forma de escoger el error lo es, de lo contrario las imputaciones serán estocásticas.

De esta fórmula se derivan las dos imputaciones explicadas anteriormente, por la media y por razón.

Estas técnicas serán más precisas cuanto mayor sea la relación entre las variables y normalmente se especifica una jerarquía en las técnicas para usar la más adecuada en cada unidad. Por ejemplo, si se ha comprobado que la técnica más precisa es la imputación por regresión y la siguiente más precisa la imputación por la media, se imputaría mediante regresión todas aquellas unidades que tengan información auxiliar disponible y para las que no lo tuvieran se usaría la imputación por la media ([Särndal y Lundström 2005](#)).

#### **Imputación por donante *Hot deck*.**

Esta técnica de imputación consiste en seleccionar a un 'donante' para la asignación del valor al receptor ([Andridge y R.J. Little 2010](#)). La selección del donante se puede hacer de diversos métodos pero el objetivo es obtener un donante lo mas similar posible para que la imputación sea mas precisa. De esta forma, una vez seleccionado el donante el proceso consistirá en imputar el valor del ítem del donante en el recipiente. Estas técnicas tienen la desventaja de que todos los donantes son parte de las observaciones y esto conlleva a asumir que no existen diferencias entre las personas que responden y las que no ([Särndal y Lundström 2005](#)).

[Andridge y R.J. Little 2010](#) definen el termino *hot deck* como el uso de un donante

disponible en el mismo conjunto de datos que los valores ausentes y se contrapone a la imputación *cold deck* que consiste en el uso de conjuntos de datos diferentes para la selección del donante como por ejemplo periodos anteriores. Entre los métodos de imputación *hot deck* más usados destacan:

- **Imputación *hot deck* aleatoria y secuencial.**

Es el método más simple de imputación usando un donante. El proceso consiste en seleccionar un donante al azar de las observaciones y usarlo para imputar un valor al recipiente. Este método tiene la ventaja de no requerir información auxiliar. En el caso de tener información adicional, como por ejemplo, la pertenencia a subgrupos de población se podría restringir la selección aleatoria del donante a ese grupo concreto. Si en vez de realizar una asignación aleatoria del donante se asigna la unidad más próxima en el registro con las características deseadas será imputación *hot deck* secuencial.

En ambos métodos la desviación típica de los totales y la media aumentará ya que siempre existe la posibilidad de que un outlier sea el donante. Las estimaciones *hot deck* serán insesgadas únicamente para cuando los valores *missing* son MCAR que es poco probable que ocurra por lo que se recomienda reducirlo utilizando la información auxiliar ([R.J.A. Little y Rubin 2002](#)).

- **Imputación por el vecino más cercano.**

Esta imputación se diferencia de las anteriores en que en vez de seleccionar una unidad con las mismas características se selecciona una unidad que minimice una función de distancia previamente definida. La función general de distancia usada es la distancia de Minkowski ([Waal, Pannekoek y Scholtus 2011](#)).

En la versión mas simple donde solamente tenemos una variable auxiliar el donante será aquel para el que la diferencia en esa variable sea mínima. En el caso de tener varias variables auxiliares el donante seria aquel que menor suma de todas las distancias tuviera pero nótese que primero las variables se deben estandarizar o usar distancias relativas como la de Mahalanobis. A cada una de estas variables se le puede aplicar una ponderación para así dar mas o menos importancia a las variables que se deseen.

La desventaja de esta técnica es que la distancia no puede calcularse de la misma forma para variables categóricas y numéricas pero esto puede solucionarse usando las variables categóricas para crear subgrupos homogéneos donde después extraer el donante ([Waal, Pannekoek y Scholtus 2011](#)).

### **Imputación por parte del experto.**

Este método de imputación es el menos estadístico ya que consiste en que el o la responsable del producto estadístico determine el valor *missing* usando su capacidad analítica y toda la información de la que disponga.

En el Tema [11](#) se verá con más detalle la imputación.

## 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.

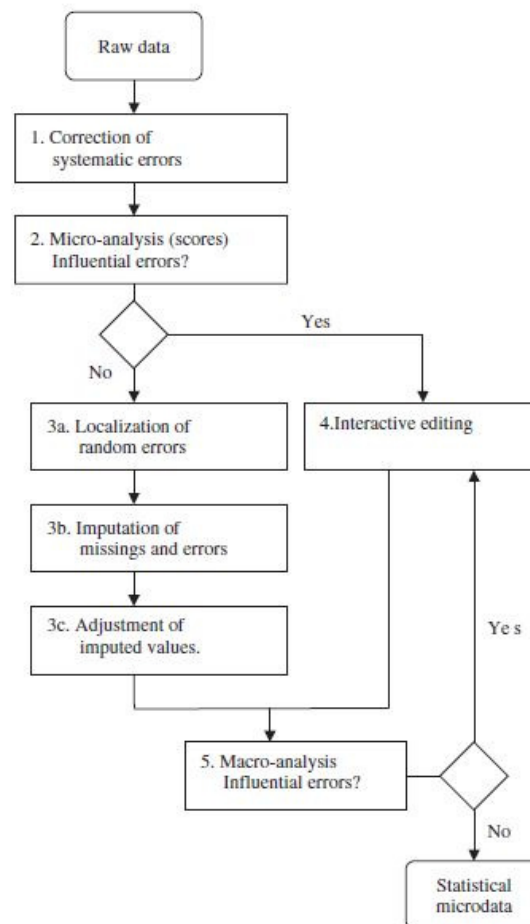


Figura 9.1: Estrategia general de depuración e imputación

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 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.

## Bibliografía

- Andridge, R.R. y R.J. Little (2010). "A review of hot deck imputation for survey non-response". En: *International Statistical Review* 78, págs. 40-64.
- 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.
- Eurostat (2014). *Handbook on Methodology of Modern Business Statistics*. URL: [https://ec.europa.eu/eurostat/cros/content/imputation\\_en](https://ec.europa.eu/eurostat/cros/content/imputation_en).
- Fellegi, I.P. y D. Holt (1976). "A systematic approach to automatic edit and imputation". En: *J. Amer. Stat. Assoc.* 71, págs. 17-35.
- 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.



- Granquist, L. (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 S. Lundström (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 28 de octubre de 2021. URL: <https://statswiki.unece.org/display/sde/GSDEM>.
- Waal, T. de, J. Pannekoek y S. Scholtus (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.
- Whitridge, P. y J. Kovar (1990). "Applications of the generalised edit and imputation system at Statistics Canada". En: *Proceedings of the Section on Survey Research Methods*, págs. 105-110.



## Tema 10

**Introducción a la estimación en presencia de falta de respuesta. Errores debidos al muestreo y a la falta de respuesta. Error cuadrático medio y sus componentes bajo falta de respuesta. Estimadores simples y sus sesgos debidos a la falta de respuesta.**

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

C.-E. Särndal y S. Lundström (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.

### 10.1 Introducción a la estimación en presencia de falta de respuesta

Supongamos que tenemos una población  $U$ , con el cual coincide exactamente el marco poblacional disponible, y queremos estudiar el total poblacional para una variable  $y$ ,  $Y = \sum_U y_k$ , y los totales de algunos dominios,  $Y_q = \sum_{U_q} y_k$ , para  $q = 1, \dots, Q$ . El diseño muestral seleccionado determina el peso de diseño  $d_k = \frac{1}{\pi_k}$  para cada elemento  $k \in U$ , donde  $\pi_k$  es la probabilidad de inclusión de  $k$ . Podemos pensar en cada variable de estudio como un apartado de un cuestionario que responde cada unidad  $k$  de la muestra  $s$ . El conjunto de respuestas de la encuesta se denota por  $r$ . Para estos elementos tenemos una respuesta por lo menos para uno de los items del cuestionario. Las unidades para las que tenemos falta de respuesta total es el conjunto complementario,  $s - r$ . La situación se puede ver en la Figura 11.1.

Para una determinada variable  $y$ , el conjunto de elementos para los cuales  $y_k$  es *missing* define el conjunto de falta de respuesta para esa variable particular. Este conjunto contiene todas las unidades con falta de respuesta total,  $s - r$ , y un conjunto adicional de elementos, aquellos con falta de respuesta parcial en la variable  $y$ .

Hay dos enfoques para tratar la falta de respuesta, *reponderación* e *imputación*. Cuando se usa la reponderación, un conjunto de pesos se fija con la ayuda de la información de una variable auxiliar, y la estimación se lleva a cabo aplicando los pesos a los valores de

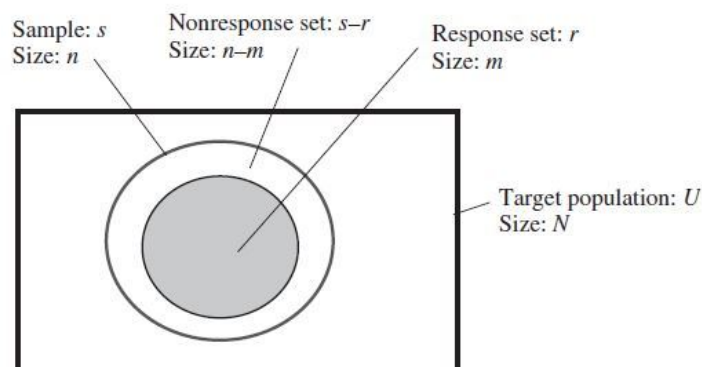


Figura 10.1: Representación de la muestra seleccionada y el conjunto de respuesta, visto como subconjuntos de la población objetivo, que se asume que es idéntica al marco poblacional.

$y$  de los elementos que han contestado. En este tema usamos el *enfoque del calibrado* para calcular los pesos.

Si  $y$  sólo se ve afectado por falta de respuesta total, y no hay falta de respuesta parcial, el estimador del parámetro de interés,  $Y = \sum_U y_k$ , será  $\hat{Y}_W = \sum_r w_k y_k$ , donde el índice  $W$  se usa para denotar 'reponderación' (del inglés *weighting*) y los pesos calibrados  $w_k$  son, por lo menos para la mayoría de los elementos, mayor que los pesos que tendría si la respuesta fuese completa (para compensar los elementos perdidos por la falta de respuesta).

El otro enfoque para el tratamiento de la falta de respuesta, la imputación, implica que para los valores  $y_k$  que son *missing* se crean valores proxy. El valor proxy creado para el elemento  $k$ , el valor imputado para  $k$ , se denota por  $\hat{y}_k$ , para distinguirlo del valor observado. Para el elemento  $k$ , la 'estimación' se lleva a cabo mediante un *método de imputación*. Hay muchos métodos de imputación y a menudo se pueden usar varios métodos en la misma encuesta.

Consideraremos que la imputación se utiliza en el caso de falta de respuesta parcial, mientras que los pesos calibrados compensa la falta de respuesta total. Para ello, crearemos primero un conjunto completo de datos consistente en valores, algunos observados, otros imputados, para cada variable y cada elemento  $k$  en el conjunto de respuestas de la encuesta  $r$ . Para una variable  $y$ , el conjunto completo de datos se puede escribir como  $\{y_{\bullet k} : k \in r\}$ , donde  $y_{\bullet k} = y_k$ , valor observado, si el elemento  $k$  responde a la variable  $y$ , e  $y_{\bullet k} = \hat{y}_k$ , valor imputado, si  $y_k$  es *missing* por falta de respuesta parcial y se imputa por  $\hat{y}_k$ .

El estimador de  $Y = \sum_U y_k$  se calcula a partir del conjunto de datos completo como  $\hat{Y}_{IW} = \sum_r w_k y_{\bullet k}$ , donde los pesos calibrados  $w_k$  compensan la falta de respuesta total. El subíndice doble  $IW$  indica 'imputación seguida de reponderación'.

## 10.2 Errores debidos al muestreo y a la falta de respuesta

Consideraremos técnicas para reducir simultáneamente el error de muestreo y el error por falta de respuesta *después de la recogida de datos*, es decir, después de que haya falta de respuesta. Estas técnicas se basan en el uso efectivo de información auxiliar. La notación general será  $\hat{Y}_W$  para el estimador calibrado e  $\hat{Y}_{IW}$  para el estimador creado por imputación seguido por reponderación por calibrado.

Incluso usando información auxiliar tenemos que aceptar que tanto  $\hat{Y}_W$  como  $\hat{Y}_{IW}$  se verán afectados por errores de muestreo y de falta de respuesta. Sin embargo, cuanto mejor sea la información auxiliar menor serán los errores.

En este tema denotaremos por  $\hat{Y}_{NR}$  el *estimador con falta de respuesta* (del inglés *non response*) tanto para  $\hat{Y}_W$  como para  $\hat{Y}_{IW}$ . Además, denotaremos por  $\hat{Y}$  la expresión de  $\hat{Y}_{NR}$  para la respuesta total, es decir, cuando  $r = s$ . Consideraremos  $\hat{Y}$  como el estimador GREG (véase el tema 4 del bloque de 'Producción Estadística Oficial: Métodos Avanzados de la especialidad de Estadística-Ciencia de datos'), el estimador HT (véase el tema 2) o algún otro estimador insesgado para repetidas muestras  $s$  obtenidas de  $U$ .

El error de  $\hat{Y}_{NR}$  es su desviación del valor objetivo  $Y = \sum_U y_k$ . Podemos escribir el error como la suma de dos componentes,

$$\hat{Y}_{NR} - Y = (\hat{Y} - Y) + (\hat{Y}_{NR} - \hat{Y}) \quad (10.1)$$

El primer término de la derecha,  $\hat{Y} - Y$ , es el *error de muestreo* (el error causado por la selección y obtención de los datos de una muestra y no del total de la población). El segundo término,  $\hat{Y}_{NR} - \hat{Y}$ , es el error por falta de respuesta (el error en que se incurre porque no se ha conseguido una respuesta completa). En una encuesta nos vamos a encontrar otros errores que no vamos a considerar aquí.

Necesitamos conocer las propiedades estadística básicas del estimador  $\hat{Y}_{NR}$ . Su *tendencia central* determinada por su valor esperado, y la *acuracidad* determinada por su error cuadrático medio. En este tema, los valores se podrán obtener sobre todas las posibles muestras  $s$  que se pueden extraer con un diseño muestral, denotadas por  $p(s)$ ; y sobre todos los posibles conjuntos de respuestas  $r$  que se pueden obtener, dada  $s$ , bajo la distribución de respuesta no conocida, denotada por  $q(r|s)$ .

Se denotará por  $\mathbb{E}_p$  y  $\mathbb{E}_q$  las esperanzas con respecto a estas dos distribuciones, respectivamente. Los operadores con respecto a ambas distribuciones conjuntamente vendrá dada por el índice  $pq$ .

La media (sobre todas las posibles muestras  $s$ ) del error de muestreo  $\hat{Y} - Y$  es cero o casi cero. La media (sobre todas las posibles muestras  $s$  y todos los posibles conjuntos

de respuesta  $r$ ) del error de falta de respuesta  $\hat{Y}_{NR} - \hat{Y}$  probablemente será distinto de cero, porque la falta de respuesta sistemáticamente produce algún sesgo.

El valor condicional (sobre  $s$ ) del error por falta de respuesta, que podemos llamar el *sesgo de falta de respuesta dada la muestra*, es

$$\mathbb{B}_{NR|s} = \mathbb{E}_q((\hat{Y}_{NR} - \hat{Y})|s) = \mathbb{E}_q(\hat{Y}_{NR}|s) - \hat{Y}. \quad (10.2)$$

Calculando la media sobre todas las posibles muestras  $s$ , obtenemos el *sesgo de falta de respuesta* (sin restricciones) de  $\hat{Y}_{NR}$ ,

$$\mathbb{B}_{NR} = \mathbb{E}_p(\mathbb{B}_{NR|s}) = \mathbb{E}_{pq}(\hat{Y}_{NR} - \hat{Y}) = \mathbb{E}_{pq}(\hat{Y}_{NR}) - \mathbb{E}_p(\hat{Y}). \quad (10.3)$$

El error de muestreo esperado, que podemos llamar *sesgo de muestreo* (del inglés *sampling bias*, por eso se usa el subíndice SAM), es

$$\mathbb{B}_{SAM} = \mathbb{E}_p(\hat{Y} - Y) = \mathbb{E}_p(\hat{Y}) - Y. \quad (10.4)$$

El sesgo (sin restricciones) de  $\hat{Y}_{NR}$ , es

$$\mathbb{B}_{pq}(\hat{Y}_{NR}) = \mathbb{E}_{pq}(\hat{Y}_{NR} - Y) = \mathbb{B}_{SAM} + \mathbb{B}_{NR}. \quad (10.5)$$

El sesgo del estimador  $\hat{Y}_{NR}$  es la suma de dos componentes, la esperanza  $p$  del error de muestreo (el sesgo de muestreo) y la esperanza  $pq$  del error de falta de respuesta (el sesgo de falta de respuesta). El primero es cero o casi nulo por razones prácticas. El sesgo de  $\hat{Y}_{NR}$  viene enteramente, o casi, del desconocido (y quizá significativo) sesgo de falta de respuesta.

El sesgo  $\mathbb{B}_{pq}(\hat{Y}_{NR})$  es en la práctica imposible de conocer, porque no se conoce nunca de forma exacta la distribución de respuesta  $q(r|s)$ . Sin embargo, en la práctica, a menudo se asume, de forma correcta o incorrecta, que el sesgo de falta de respuesta es 'suficientemente pequeño'. A menudo, esta hipótesis no está justificada.

En relación con la varianza,  $\mathbb{V}_{pq}(\hat{Y}_{NR}) = \mathbb{E}_{pq}(\hat{Y}_{NR} - \mathbb{E}_{pq}(\hat{Y}_{NR}))^2$ , se puede escribir como la suma de dos componentes,

Varianza = Varianza muestral + Varianza de falta de respuesta.

De forma más específica, si el sesgo condicional  $\mathbb{B}_{NR|s}$  es nulo para cada muestra posible  $s$ , entonces

$$V_{pq}(\hat{Y}_{NR}) = V_{SAM} + V_{NR}. \quad (10.6)$$

donde  $V_{SAM} = \mathbb{V}_p(\hat{Y})$  se llama *varianza muestral*, y  $V_{NR} = \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s)$  se llama *varianza de falta de respuesta*.

La componente  $V_{SAM}$  es la varianza de  $\hat{Y}$  sobre todas las posibles muestras  $s$  que se pueden obtener a partir de un diseño muestral; no depende de la falta de respuesta ni de la distribución de respuesta. La componente  $V_{NR}$  implica calcular la media sobre todas las muestras  $s$  así como sobre todos los conjuntos de respuesta  $r$ .

La condición de que  $\mathbb{B}_{NR|s} = 0$  para todas las posibles muestras  $s$  es muy fuerte. Si no se verifica, se añadirán más componentes a la varianza de falta de respuesta, por lo que  $V_{NR}$  en (10.6) se convierte en

$$V_{NR} = \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s) + \mathbb{V}_p(\mathbb{B}_{NR|s}) + 2\mathbb{C}_p(\hat{Y}, \mathbb{B}_{NR|s}). \quad (10.7)$$

La ecuación (10.7) indica que incluso en las circunstancias favorables en que la falta de respuesta no cause sesgo, el efecto será un aumento en la varianza, comparado con la respuesta total, en cuyo caso sólo tendríamos la componente  $V_{SAM} = \mathbb{V}_p(\hat{Y})$ . Más aún, si hay sesgo, la varianza sufre más inflación, a través de los dos últimos términos del lado derecho de (10.7).

En la práctica, lo que podemos esperar es que el procedimiento de estimación tendrá éxito en la reducción del sesgo a niveles bajos, de forma que  $\mathbb{B}_{NR|s} = 0$  se verifique al menos aproximadamente para cada muestra  $s$ .

Es decir, nos interesará evaluar la varianza de falta de respuesta  $V_{NR}$ , ya que si nos encontramos de forma regular con que significa un porcentaje muy alto de la varianza total, es una señal clara de que es necesario reducir la falta de respuesta en el futuro.

### 10.3 Error cuadrático medio y sus componentes bajo falta de respuesta

La expresión de la varianza para  $\hat{Y}_{NR}$  en (10.6) se obtiene de la siguiente forma. Hay dos fases de selección.  $s$  se selecciona a partir de  $U$ , y, dado  $s$ ,  $r$  se obtiene como un subconjunto de  $s$ . Las dos distribuciones de probabilidad son  $p(s)$  y  $q(r|s)$ . Bajo estas circunstancias, podemos obtener la varianza de una cantidad aleatoria por la regla de 'la varianza de la esperanza condicional más la esperanza de la varianza condicional'. Por (10.2), la esperanza condicional es  $\mathbb{E}_q(\hat{Y}_{NR}|s) = \hat{Y} + \mathbb{B}_{NR|s}$ . Obtenemos la varianza  $pq$

$$\mathbb{V}_{pq}(\hat{Y}_{NR}) = \mathbb{V}_p(\hat{Y} + \mathbb{B}_{NR|s}) + \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s). \quad (10.8)$$

Si  $\mathbb{B}_{NR|s} = 0$  para cada  $s$ , obtenemos el resultado (10.6) con  $V_{NR} = \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s)$ . Si esta condición no se verifica, entonces extendemos el término  $\mathbb{V}_p(\hat{Y} + \mathbb{B}_{NR|s})$  en (10.8) para encontrar el resultado (10.6) con  $V_{NR}$  dado por (10.7).

En la presencia de sesgo, un indicador de variabilidad más relevante que la varianza es el error cuadrático medio (ECM, MSE en inglés). El MSE  $pq$  de  $\hat{Y}_{NR}$ ,  $MSE_{pq}(\hat{Y}_{NR})$ ,

es la media del error cuadrático,  $(\hat{Y}_{NR} - Y)^2$ , sobre todas las muestras  $s$  y todos los conjuntos de respuesta  $r$  contenidos en  $s$ . Usando el hecho de que  $MSE_{pq}(\hat{Y}_{NR}) = \mathbb{V}_{pq}(\hat{Y}_{NR}) + (\mathbb{B}_{pq}(\hat{Y}_{NR}))^2$ , obtenemos, después de simplificar,

$$MSE_{pq}(\hat{Y}_{NR}) = \mathbb{V}_p(\hat{Y}) + \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s) + \mathbb{E}_p(\mathbb{B}_{NR|s}^2) + 2\mathbb{C}_p(\hat{Y}, \mathbb{B}_{NR|s}) + 2B_{SAM}B_{NR} + (B_{SAM})^2. \quad (10.9)$$

Aquí,  $\mathbb{V}_p(\hat{Y})$  es la varianza del estimador de respuesta completa. Los términos principales del lado derecho son los tres primeros. Por tanto, podemos escribir:

$$MSE_{pq}(\hat{Y}_{NR}) \approx \mathbb{V}_p(\hat{Y}) + \mathbb{E}_p \mathbb{V}_q(\hat{Y}_{NR}|s) + \mathbb{E}_p(\mathbb{B}_{NR|s}^2). \quad (10.10)$$

Aquí el término  $\mathbb{E}_p(\mathbb{B}_{NR|s}^2)$ , provocado por el sesgo de falta de respuesta, puede representar un aumento considerable del MSE.

## 10.4 Estimadores simples y sus sesgos debidos a la falta de respuesta

Ante un procedimiento de obtención de información, como las encuestas, en el que hay falta de respuesta o éstas no se ajustan a las preguntas, se pueden 'construir' estimadores específicos particulares para el tratamiento de las mismas. Para ello se tendrá muy en cuenta el alcance de la información auxiliar existente.

El enfoque general está orientado a tomar de base una amplia familia de estimadores  $\hat{Y}_W$ , tal que sus miembros corresponden a diferentes fuentes de información. En la práctica, los procedimientos para hallar las fórmulas para los casos especiales de  $\hat{Y}_W$  se centrarán en cómo definir un vector auxiliar  $X_k$  adecuado.

Por ello, en este apartado se analizan algunos casos especiales de todos los estimadores de calibración posibles.

Sean las especificaciones simples del vector auxiliar y la entrada de información asociada, se obtiene el parámetro explícito de  $\hat{Y}_W$ , utilizando como norma la ponderación estándar vista en la Sección 10.1, denotando a los pesos calibrados por  $w_k$ . Sólo en algunos casos requiere una ponderación más general.

Además, el muestreo de la población  $U$  de dimensión  $N$  se realiza con un diseño muestral con ponderaciones muestrales tales que  $d_k = \frac{1}{\pi_k}$ . De esta forma, obteniendo  $\hat{Y}_W$  para un diseño general, es sencillo encontrar su expresión para diseños particulares.

Por motivos prácticos en la formulación, consideramos:

- un muestreo aleatorio simple (m.a.s.) de  $n$  elementos escogidos sobre un total de  $N$ , tal que  $d_k = \frac{N}{n}$  para todo  $k \in U$ .

- un muestreo aleatorio simple estratificado siendo las ponderaciones  $d_k = \frac{N_h}{n_h}$  para todo elemento  $k$  del estrato  $h$ .

De esta forma, se comienza con el vector auxiliar más sencillo que puede ir aumentando en complejidad en función de las necesidades. Al respecto, se tienen en cuenta diferentes características técnicas del vector auxiliar  $x_k = x^*_k$  en el InfoU (nivel de información de la población U) y del vector auxiliar  $x_k = x^o_k$  en el InfoS (nivel de información de la muestra s).

### 10.4.1 Vector auxiliar

El vector auxiliar más sencillo, cuyo valor es constante para todo  $k$ , no reconoce diferencias entre elementos y es ineficaz para solventar la falta de respuesta. Para el InfoU con  $x_k = x^*_k = 1$  para todo  $k$ , el estimador de calibración es

$$\hat{Y}_W = N\bar{y}_{r;d} = \hat{Y}_{EXP} \quad (10.11)$$

donde  $\bar{y}_{r;d} = \frac{\sum_r d_k y_k}{\sum_r y_k}$ . El subíndice EXP refleja el estimador de expansión (directo) a menudo usado y que en la literatura, se suele reservar más específicamente para la forma (10.11). En ese caso, con  $d_k = \frac{N}{n}$  para todo  $k$ , tenemos:

$$\hat{Y}_{EXP} = \frac{N}{m} \sum_r y_k = N\bar{y}_r \quad (10.12)$$

El sesgo de  $\hat{Y}_{EXP}$  puede ser muy grande, debido a la debilidad de la información auxiliar. De forma excepcional,  $\hat{Y}_{EXP}$  puede ser utilizado si no existe mejor información auxiliar y/o si hay una razón de peso para creer que la falta de respuesta es fruto del azar. Es posible obtener en algún momento un valor para  $\hat{Y}_{EXP}$  a partir de una encuesta, pero tan sólo para proporcionar una estimación de referencia con la que comparar los resultados de alternativas mejores.

La media  $\bar{y}_{r;d}$  sirve de ejemplo en un sistema de notación que puede ser utilizado con varias cantidades, siendo la media en este caso una media ponderada. La regla general es que el primer índice denota el conjunto sobre el que se calcula la cantidad, tal que si es U, esa cantidad no es aleatoria; pero si es igual a  $r$  o  $s$ , la cantidad sí es aleatoria.

Por tanto, en el caso de  $\bar{y}_{r;d} = \frac{\sum_r d_k y_k}{\sum_r y_k}$ , los cálculos son realizados con los elementos del conjunto respuesta  $r$ , usando  $d_k = \frac{1}{\pi_k}$  como ponderaciones. La ausencia del segundo índice implica que la ponderación es uniforme, como en  $\bar{y}_r = \frac{1}{m} \sum_r y_k$ .

### 10.4.2 Clasificación unidireccional

Ante la falta de respuesta, la información disponible sobre clasificación de elementos es muy útil para las estimaciones, de tal forma que sea factible asignar un elemento  $k$  en un

conjunto de  $P$  categorías, siendo éstas grupos mutuamente excluyentes y exhaustivos. Por ejemplo, grupos de edad, o grupos definidos por clasificación cruzada (grupo de edad por sexo, por grupo ocupacional). El identificador de grupo para el elemento  $k$  se define como:

$$\gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})' \quad (10.13)$$

donde, para  $p = 1, \dots, P$ , tenemos que  $\gamma_{pk} = 1$  si  $k$  pertenece al grupo  $p$ , y tenemos que  $\gamma_{pk} = 0$  si no pertenece. Es decir, el vector  $\gamma_k$  tiene  $P - 1$  componentes cero, y por tanto, un único componente que identifica la pertenencia o no al grupo del elemento  $k$ . La información sobre  $\gamma_k$  correspondiente a InfoU y la información sobre  $\gamma_k$  correspondiente a InfoS representa dos casos diferentes, si bien, nosotros consideramos ambos casos. La condición dada por  $\mu' \mathbf{x}_k = 1$  para todo  $k$  se cumple tomando  $\mu' = (1, 1, \dots, 1)$ . De esta manera, se aplica la fórmula:

$$w_k = w_{Mk} = \mathbf{X}' \left( \sum_r d_{\alpha k} \mathbf{z}_k \mathbf{x}_k' \right)^{-1} d_{\alpha k} \mathbf{z}_k \quad (10.14)$$

En el caso de InfoU, el vector auxiliar es  $\mathbf{x}_k = \mathbf{x}_k^* = \gamma_k$  teniendo la información asociada  $\sum_U \mathbf{x}_k^* = (N_1, \dots, N_p, \dots, N_P)'$ , donde  $N_p$  es el tamaño conocido del grupo  $U_p$ . Esto se aplica en muchas encuestas, particularmente en los países escandinavos, donde los registros disponibles de la población total pueden generar una variedad de clasificaciones para los individuos en función de sus características demográficas, educativas, económicas y otras características personales. Las frecuencias de población  $N_p$  son establecidas por un simple conteo en el registro. En otros países, el valor  $N_p$  puede definirse a partir de conteos censales (actualizados).

Sea  $r_p$  el conjunto respuesta del grupo  $p$ , entonces el conjunto respuesta total es  $r = \bigcup_{p=1}^P r_p$ . Partiendo de la fórmula estándar ponderada:

$$w_k = d_k v_k, \quad v_k = 1 + \lambda_r' \mathbf{x}_k \quad (10.15)$$

teniendo que  $\lambda_r' = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}$

O también, considerando la ecuación anterior (10.14) con  $d_{\alpha k} = d_k$  y  $\mathbf{z}_k = \mathbf{x}_k$ , obtenemos  $v_k = F_p^*$  para todo  $k \in r_p$ , donde  $F_p^* = N_p / \sum_{r_p} d_k$ . Así, el estimador de calibración  $\hat{Y}_W = \sum_r w_k y_k$  entonces es:

$$\hat{Y}_W = \sum_{p=1}^P N_p \bar{y}_{r_p; d} = \hat{Y}_{\text{PWA}} \quad (10.16)$$



donde  $\bar{y}_{r_p;d} = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k}$  es la media ponderada de grupo diseñada para los encuestados. En caso de falta de respuesta por azar, dentro de cada grupo, entonces  $\bar{y}_{r_p;d}$  es el estimador de la media de grupo  $\bar{y}_{U_p} = \frac{\sum_{U_p} y_k}{N_p}$  esencialmente sin sesgo, y entonces (10.16) es casi sin sesgo para  $Y = \sum_U y_k$ .

En particular, para m.a.s e InfoU, la ecuación (10.16) se convierte así en:

$$\hat{Y}_{\text{PWA}} = \sum_{p=1}^P N_p \bar{y}_{r_p} \quad (10.17)$$

donde  $\bar{y}_{r_p} = \frac{\sum_{r_p} y_k}{m_p}$  es la estricta media de  $y$  para los  $m_p$  encuestados en el grupo  $p$ . La anterior fórmula (10.17) en ocasiones se denomina estimador post-estratificado, si bien, esta denominación puede dar lugar a confusión ya que el estimador post-estratificado tradicional se refiere a una sola fase de muestreo. Por tanto, autores como [Kalton y Kasprzyk 1986](#) distinguen entre el estimador post-estratificado, utilizado para la respuesta completa en el muestreo monofásico, y el indicado en (10.17), que denominan estimador de ajuste de ponderación poblacional, como se refleja en la notación  $\hat{Y}_{\text{PWA}}$ .

Así, en éste último estimador, se reconoce una fase previa de muestreo seguida de una fase de falta de respuesta. A lo largo de los años, varios autores han discutido (10.17), incluidos [Bethlehem y Kersten 1985](#) y [Thomsen 1973; Thomsen 1978](#). Pero, por importante que sea (10.17), es sólo una de las aplicaciones más simples del método de calibración mucho más general.

Consideremos ahora InfoS con el vector auxiliar  $\mathbf{x}_k = \mathbf{x}_k^o = \gamma_k$ . En este caso, la información disponible establece que, para cada  $k \in s$  (y así para cada  $k \in r$ ), podemos asignar  $k$  al grupo al que pertenece. Además, sea  $s_p$ , de tamaño  $n_p$ , la parte de la muestra  $s$  del grupo  $p$ ; siendo  $s = \bigcup_{p=1}^P s_p$ . De esta forma, tenemos ejemplos de situaciones en las que se puede establecer una clasificación para cada  $k \in s$  (pero no necesariamente para todo  $k \in U$ ):

- (i) Cada elemento de la muestra puede atribuirse a uno de los  $P$  posibles entrevistadores responsables de la recogida de datos.
- (ii) Las categorías representan  $P$  diferentes grupos socioeconómicos para los que los tamaños poblacionales no están disponibles para ser usados en la estimación correspondiente.
- (iii) En una encuesta telefónica con repetidos intentos de contacto, los elementos  $k \in s$  se agrupan por el intento (primero, segundo, ...,  $P$ -ésimo) en el que se establece el contacto.

En el apartado (iii), el hecho de contactar telefónicamente con alguna unidad para participar en proceso de encuesta puede suponer la aceptación o el rechazo, y eso implica

la falta de respuesta. La evidencia empírica sugiere que la tasa de participación  $\frac{m_p}{n_p}$  es mayor entre las personas que son contactados inicialmente, es decir, para  $p = 1$  o  $2$ , que para, digamos,  $p = 5$  o  $6$ . Los grupos están, por tanto, relacionados con la propensión a responder. Para el grupo final,  $n_P$  es el número de personas en  $s$  que hasta ahora no se han alcanzado, aunque algunos podrían haberse alcanzado si el procedimiento hubiera continuado.

Aunque se desconoce el valor de  $N_p$ , hay suficiente información para calcular  $\hat{N}_p = \sum_{s_p} d_k$  para  $p = 1, \dots, P$ . De la anterior ecuación (10.15) se obtiene  $v_k = F_p^\circ$  para todo  $k \in r_p$ , donde  $F_p^\circ = \hat{N}_p / \sum_{r_p} d_k$ . El estimador de calibración  $\hat{Y}_W = \sum_r w_k y_k$  queda de la forma:

$$\hat{Y}_W = \sum_{p=1}^P \hat{N}_p \bar{y}_{r_p;d} = \hat{Y}_{WC} \quad (10.18)$$

Partiendo de una m.a.s.  $n$  de una población de tamaño  $N$ , e InfoS, de los  $n$  elementos muestreados, sea  $n_p$  el número de ocurrencias del grupo  $p$ , de los que  $m_p$  responden. Las ponderaciones obtenidas de (10.15) son  $w_k = \frac{N n_p}{n m_p}$ , las cuales suponen una 'expansión por grupos, utilizando la tasa de respuesta de grupo inversa'. Entonces (10.18) queda como sigue:

$$\hat{Y}_{WC} = \sum_{p=1}^P \hat{N}_p \bar{y}_{r_p}, \quad \text{con} \quad \hat{N}_p = N n_p / n. \quad (10.19)$$

Este estimador es conocido como el estimador de clase ponderado, y se denota  $\hat{Y}_{WC}$ , habiendo sido discutido a menudo en la literatura, por ejemplo, en [Oh y Scheuren 1983](#); [Kalton y Kasprzyk 1986](#); [Little 1986](#) así como en [Sweden 1980](#).

Si bien puede parecer sorprendente al principio, (10.16) y (10.18) tienen en esencia el mismo sesgo por falta de respuesta. La diferencia de sesgo es insignificante, aunque  $\hat{Y}_{PWA}$  utiliza información a un nivel superior a  $\hat{Y}_{WC}$ . Esto se refleja de forma evidente, ya que (10.16) normalmente tiene la varianza más pequeña de las dos.

El muestreo aleatorio simple estratificado juega un papel destacado en la realización de encuestas y merece una atención especial. Se parte de muestreo aleatorio simple estratificado e InfoU, de tal manera que cada estrato también se toma como un grupo de cara al proceso de calibración. Entonces, el índice  $p$  en el InfoU vector  $\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$  indica un  $P$  posible estrato. Siendo las ponderaciones muestrales  $d_k = \frac{N_p}{n_p}$ , la inversa de la tasa de muestreo en el estrato  $p$ , se tiene  $v_k = F_p^* = \frac{n_p}{m_p}$  para todo  $k \in r_p$ , llevando a  $w_k = d_k v_k = \left(\frac{N_p}{n_p}\right) \left(\frac{n_p}{m_p}\right) = \frac{N_p}{m_p}$ . Entonces (10.16) se convierte en:

$$\hat{Y}_W = \sum_{p=1}^P N_p \bar{y}_{r_p} \quad (10.20)$$

Aunque (10.17) sea idéntica en forma, la expresión (10.20) es conceptualmente diferente siendo cada grupo también un estrato en términos de muestreo. Por tanto, el enfoque de calibración proporciona una justificación a nivel práctico con un muestreo aleatorio simple estratificado, para la expansión de las ponderaciones  $\frac{N_p}{n_p}$  estrato por estrato, utilizando la inversa de la tasa de respuesta del estrato  $\frac{n_p}{m_p}$ . Así, para muestreo aleatorio simple estratificado é InfoS, tenemos  $\mathbf{x}_k = \mathbf{x}_k^o = \gamma_k$ , y llegamos de nuevo a la expresión (10.20).

Algunos autores son de la idea de que primero deberíamos ajustar las ponderaciones, estrato por estrato, y luego aplicar una calibración. Es decir, antes de la calibración se ajustarían los pesos muestrales dentro de cada estrato, usando el inverso de la tasa de respuesta del estrato,  $\frac{n_p}{m_p}$ , para que los pesos iniciales sean  $d_{\alpha k} = \left(\frac{N_p}{n_p}\right) \left(\frac{n_p}{m_p}\right) = \frac{N_p}{m_p}$  para  $k \in r_p$ .

Sin embargo, esto supone cambiar las ponderaciones muestrales, estrato por estrato, y no por un factor que sea constante en todo momento. Usando (10.15) con los pesos iniciales  $d_{\alpha k} = \frac{N_p}{m_p}$  y  $\mathbf{x}_k = \mathbf{x}_k^* = \mathbf{z}_k = \gamma_k$ , obtenemos  $v_k = 1$  para todo  $k$ . Las ponderaciones finales son exactamente iguales que en (10.20), es decir,  $w_k = \frac{N_p}{m_p}$  para  $k \in r_p$ . Resumiendo, los pasos realizados en la calibración no han supuesto ningún cambio, ya que los pesos iniciales y los finales coinciden.

Sin embargo, en muestreo aleatorio simple estratificado no es cierto de manera general que las dos alternativas consideradas para las ponderaciones,  $d_{\alpha k} = d_k = \frac{N_p}{n_p}$  y  $d_{\alpha k} = \frac{N_p}{m_p}$ , den idénticos pesos finales, depende del vector  $\mathbf{x}_k$ . Es probable que la diferencia entre las dos estimaciones calibradas no tenga consecuencias en la mayoría de los casos, pero la cuestión merece prestar especial atención.

### 10.4.3 Una sola variable auxiliar cuantitativa

El siguiente ejemplo es poco común ya que alcanza el objetivo de llegar a una fórmula bien conocida utilizando un vector instrumental  $\mathbf{z}_k$  en lugar de usar la opción estándar  $\mathbf{z}_k = \mathbf{x}_k$ . Se considera una variable auxiliar cuantitativa,  $x$ , con valor  $x_k$  para el elemento  $k$ . Puede ser una medida del tamaño de  $k$ , como el número de empleados de una empresa  $k$  tomados en una encuesta económica. Para cada  $k \in r$  se obtiene el valor  $x_k$  o se conoce por otros medios. De tal forma que surgen dos posibilidades: utilizar solo  $x_k$ , o, suponiendo que el valor  $N$  es conocido, utilizar el vector  $(1, x_k)'$ . Tendremos en cuenta ambas posibilidades.

Considerando InfoU con  $\mathbf{x}_k = \mathbf{x}_k^* = x_k$  y el correspondientes valor total poblacional  $\sum_U x_k$ , partiendo de (10.15) con  $d_{\alpha k} = d_k$  y  $\mathbf{z}_k = 1$  para todo  $k$ , se obtiene:

$$\hat{Y}_W = \left( \sum_U x_k \right) \frac{\sum_r d_k y_k}{\sum_r d_k x_k} = \hat{Y}_{RA} \quad (10.21)$$

Esta ecuación es del tipo de *estimador de razón*, de ahí el índice. Si la falta de respuesta ocurre con la misma probabilidad en toda la población, entonces (10.21) es casi sin sesgo para  $Y = \sum_U y_k$

La función del estimador de razón se vuelve aún más explícita bajo m.a.s., en cuyo caso

$$\hat{Y}_{RA} = N \bar{x}_U \frac{\bar{y}_r}{\bar{x}_r} \quad (10.22)$$

donde  $\bar{x}_U = \frac{\sum_U x_k}{N}$ ,  $\bar{y}_r = \frac{\sum_r y_k}{m}$ , y  $\bar{x}_r$  se definen análogamente. Hacer notar que el estimador de razón sobre el que normalmente se discute en los libros corresponde a (10.21), la cual en m.a.s. es casi imparcial. Sin embargo, esta propiedad no ocurre para (10.22) a menos que la falta de respuesta ocurra completamente al azar.

En el caso de InfoS, conocemos el valor  $x_k$  para cada  $k \in s$ . Obtenemos las ponderaciones de (10.15) con  $\mathbf{x}_k = \mathbf{x}_k^o = x_k$ ,  $d_{\alpha k} = d_k$  y  $\mathbf{z}_k = 1$  para todo  $k$ . En particular, teniendo m.a.s. é InfoS, la muestra  $x$ -media  $\bar{x}_s = \frac{\sum_s x_k}{n}$  se puede calcular, y el estimador es:

$$\hat{Y}_W = N \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r} \quad (10.23)$$

En presencia de una variable continua  $x$ , se puede formular de manera alternativa el vector auxiliar como  $\mathbf{x}_k = (1, x_k)'$ . Para InfoU, tenemos  $\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)'$  con la información de entrada  $\mathbf{X} = \sum_U \mathbf{x}_k = (N, \sum_U x_k)'$ , asumiendo que es conocido el tamaño de la población  $N$ . Así, con las ponderaciones estándar (10.15) se tiene que:

$$\hat{Y}_W = N \{ \bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d}) B_{r;d} \} = \hat{Y}_{REG} \quad (10.24)$$

que es la forma del estimador de regresión, donde  $\bar{x}_{r;d} = \frac{\sum_r d_k x_k}{\sum_r d_k}$  y

$$B_{r;d} = \frac{\sum_r d_k (x_k - \bar{x}_{r;d}) (y_k - \bar{y}_{r;d})}{\sum_r d_k (x_k - \bar{x}_{r;d})^2} \quad (10.25)$$

Las bondades de  $\hat{Y}_{\text{REG}}$  como estimador ajustado a la falta de respuesta son discutidas, por ejemplo, en [Bethlehem 1988](#). El estimador de regresión simple clásico, analizado en muchos libros de texto para el caso de respuesta completa, se obtiene de (10.24) cuando  $r = s$ . Más aún, el estimador (10.24) normalmente resulta mejor ante el sesgo por falta de respuesta que el estimador de razón (10.21).

El estimador de razón tiene una larga historia de excelente desempeño en la teoría de encuestas, siempre que no haya falta de respuesta. Pero, extender su utilización en encuestas con falta de respuestas es arriesgado. No es recomendable (10.21) o (10.22). Para que estos estimadores sean casi insesgados, las probabilidades de respuesta deben satisfacer una condición muy fuerte: deben ser iguales en toda la población.

Por lo tanto, el estimador de razón está en desventaja en comparación con alternativas como el estimador de clasificación unidireccional visto. Este último es casi insesgado con una probabilidad de respuesta igual dentro de los grupos, algo que es más fácil de satisfacer que la probabilidad de respuesta igual en toda la población. Esta observación implica un uso modificado y de más confianza de una variable cuantitativa  $X$ :  $P$ -grupos definidos por intervalos no superpuestos de los valores  $x_k$  disponibles para  $k \in U$  o para  $k \in s$ ; entonces el vector  $x_k$  se convierte en un vector de clasificación

$$\gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$$

Alternativamente, se podría formular  $x_k$  tal que:

$$x_k = x_k^* = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk}, \gamma_{1k}x_k, \dots, \gamma_{pk}x_k, \dots, \gamma_{Pk}x_k)'$$

#### 10.4.4 Clasificación unidireccional combinada con un variable cuantitativa

En este caso, la información auxiliar está presente tanto sobre una variable categórica de valor  $P$ , como sobre una variable cuantitativa  $x$  que puede medir el tamaño de un cierto elemento. La información del elemento respuesta  $k$  es el valor observado  $x_k$ , y el grupo al que pertenece codificado por el indicador de grupo  $\gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$  definido en el apartado anterior. De nuevo surgen dos casos para formar el vector auxiliar: i), partiendo del mismo vector  $x_k$ ; ii) alternativamente, con el vector  $(1, x_k)'$ .

i) Para InfoU, tenemos  $x_k = x_k^* = (\gamma_{1k}x_k, \dots, \gamma_{pk}x_k, \dots, \gamma_{Pk}x_k)'$ . La entrada de información requerida,  $\sum_U x_k^*$ , es el vector compuesto por los totales del grupo  $P$ , tal que  $\sum_U x_k$ . Más adelante, en esta sección, se considera una formulación de  $x_k$  que también tiene en cuenta los tamaños conocidos de los grupos  $N_p$ . Sin embargo, de (10.15) con  $x_k = \gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$  se obtiene otra forma de estimador bien conocida:

$$\hat{Y}_W = \sum_{p=1}^P \left( \sum_{U_p} x_k \right) \frac{\bar{y}_{r_p;d}}{\bar{x}_{r_p;d}} = \hat{Y}_{\text{SEPREA}} \quad (10.26)$$

donde  $\bar{y}_{r_p;d}$  se definió en el anterior apartado de *Clasificación unidireccional* y  $\bar{x}_{r_p;d}$  es análoga, siendo  $\hat{Y}_{\text{SEPREA}}$  (del inglés *separated ratio estimator*) un *estimador de razón separado* construido como una suma de estimadores de razón, uno para cada grupo.

ii) Para InfoS, tenemos  $x_k = x_k^o = (\gamma_{1k}x_k, \dots, \gamma_{pk}x_k, \dots, \gamma_{Pk}x_k)'$ . Partiendo de (10.15) con  $d_k = \frac{N}{n}$  y  $x_k = \gamma_k$  se obtiene  $w_k = \frac{\sum_{s_p} d_k x_k}{\sum_{r_p} d_k x_k}$  para todo  $k$  del grupo  $p$ . El resulta, de nuevo, es un estimador fácilmente reconocible:

$$\hat{Y}_W = \sum_{p=1}^P \left( \sum_{s_p} d_k x_k \right) \frac{\bar{y}_{r_p;d}}{\bar{x}_{r_p;d}} \quad (10.27)$$

En comparación con  $\hat{Y}_{\text{RA}}$  dado en (10.21), tanto (10.26) como (10.27) ofrecen mejor protección contra sesgo por falta de respuesta.

Para incorporar información que puede existir sobre los tamaños de los grupos de población  $N_p$ , formulamos en su lugar el vector auxiliar para InfoU como

$$x_k = x_k^* = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk}, \gamma_{1k}x_k, \dots, \gamma_{pk}x_k, \dots, \gamma_{Pk}x_k)' \quad (10.28)$$

Tomando las ponderaciones obtenidas por (10.15), el estimador resultante es:

$$\hat{Y}_W = \sum_{p=1}^P N_p \{ \bar{y}_{r_p;d} + (\bar{x}_{U_p} - \bar{x}_{r_p;d}) B_{r_p;d} \} = \hat{Y}_{\text{SEPREG}} \quad (10.29)$$

donde  $\bar{x}_{U_p} = \sum_{U_p} \frac{x_k}{N_p}$ ,  $\bar{y}_{r_p;d}$ ,  $\bar{x}_{r_p;d}$  ya se definieron, y  $B_{r_p;d}$  viene dado por (10.25) si reemplazamos  $r$  por  $r_p$ . La notación  $\hat{Y}_{\text{SEPREG}}$  sugiere otra forma bien conocida, la del *estimador de regresión separado*. De los estimadores vistos hasta ahora en este tema, SEPREG tiene la mejor protección efectiva contra el sesgo por falta de respuesta.

### 10.4.5 Clasificación bidireccional

En la práctica, a menudo existe información sobre dos o más variables auxiliares categóricas o factores. En este apartado, consideraremos el caso de dos factores. El razonamiento puede ser fácilmente extendido a una clasificación multifactorial.

- Para el primer factor, supongamos, como en el caso unidireccional, que hay  $P$  categorías  $p = 1, \dots, P$ , que representan, por ejemplo, una clasificación geográfica.
- Para el segundo factor, sean categorías  $H$  indexadas  $h = 1, \dots, H$ , tal que representa, por ejemplo, una clasificación socioeconómica.

Entonces se puede colocar un elemento respuesta  $k$  en una de las celdas  $P \times H$  formada por la clasificación cruzada de los dos factores.

Si se sigue el procedimiento de *clasificación unidireccional* formulamos un vector auxiliar con dimensión  $P \times H$ . Para InfoU, esto requiere los valores de las celdas  $P \times H$  de población  $N_{ph}$  se conozcan. Pero para otras formulaciones, más simples, es posible que sea necesario conocer el vector auxiliar.

Considerando el vector indicador siguiente:

$$(\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk}, \delta_{1k}, \dots, \delta_{hk}, \dots, \delta_{Hk})' \quad (10.30)$$

donde los  $\gamma$ s indican la primera clasificación con  $P$  grupos y los  $\delta$ s indican la segunda clasificación con  $H$  grupos. Es decir,  $\gamma_{pk}$  es como en la sección de *clasificación unidireccional*, mientras que  $\delta_{hk} = 1$  si  $k$  es un miembro del grupo  $h$ , y  $\delta_{hk} = 0$  si no, para  $h = 1, \dots, H$ .

Se debe formular el vector auxiliar  $x_k$  para evitar una singularidad de la matriz que debe invertirse cuando se calculan los pesos calibrados. Existe una dependencia lineal en el vector (10.30): para cada  $k$  tenemos  $\sum_{p=1}^P \gamma_{pk} = \sum_{h=1}^H \delta_{hk} = 1$ , y la matriz a invertir es singular. Por lo tanto, se elimina cualquiera de los  $\delta$ s en la segunda clasificación, digamos, la última. Entonces, para InfoU, el vector auxiliar es:

$$x_k = x_k^* = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk}, \delta_{1k}, \dots, \delta_{hk}, \dots, \delta_{H-1,k})' \quad (10.31)$$

Su dimensión es  $P+H-1$  y la correspondiente información de entrada es  $\sum_U x_k = \sum_U x_k^*$ , es un vector con  $P+H-1$  marginales, tal que  $N_{p\bullet} = \sum_{h=1}^H N_{ph}$ , con  $p=1, \dots, P$ , y  $N_{\bullet h} = \sum_{p=1}^P N_{ph}$ , con  $h=1, \dots, H-1$ , donde  $N_{ph}$  es el total poblacional en la celda  $ph$ . De este modo obtenemos el procedimiento conocido como 'calibración en conteos marginales'. La formulación del vector auxiliar (10.31) nos permite tratar tres situaciones que ocurren comúnmente:

- (i) Los  $P \times H$  conteos de celdas y  $N_{ph}$ ,  $p = 1, \dots, P$ ,  $h = 1, \dots, H$ , son conocidos, pero se considera que los conteos marginales,  $N_{p\bullet}$ ,  $p = 1, \dots, P$ , y  $N_{\bullet h}$ ,  $h = 1, \dots, H-1$ ,



contienen casi la misma cantidad de información, por lo que en su lugar usamos (10.31), sin incurrir en pérdidas graves.

(ii) Los  $P \times H$  conteos de celdas y  $N_{ph}, p = 1, \dots, P, h = 1, \dots, H$ , son conocidos, pero algunos de ellos son extremadamente pequeños o nulos, situación que surge con frecuencia en la práctica. 'Colapsar' las celdas es un remedio alternativo para este problema.

(iii) Los conteos marginales, tal que  $N_{p\bullet}$  y  $N_{\bullet h}$ , son conocidos, pero no así los conteos de las celdas  $N_{ph}$ . Un ejemplo importante de esto ocurre cuando  $N_{p\bullet}$  y los  $N_{\bullet h}$  son conteos derivados de dos registros diferentes. No es necesario combinarlos en función de los elementos básicos, y esto puede ser una ventaja considerable. Existen dos posibilidades para preparar el escenario InfoU: (a) obtener miembros de clase para  $k \in r$ , tal que el conjunto de respuestas es emparejado usando una clave única, como el número de identificación personal, con el primer registro, luego con el segundo, mientras que  $N_{p\bullet}$  y  $N_{\bullet h}$  se establecen por conteo en cada registro de forma separada; (b) la información de los miembros de clase se solicita en la encuesta y se obtiene para cada  $k \in r$ , mientras los conteos son importados (obtenidos),  $N_{p\bullet}$  para  $p = 1, \dots, P$  de un registro y  $N_{\bullet h}$  para  $h = 1, \dots, H$ , de otro.

Para el vector auxiliar (10.31), las ponderaciones estándar  $w_k$  obtenidas de (10.15) dan el estimador de calibración:

$$\hat{Y}_W = \sum_r w_k y_k = \hat{Y}_{\text{TWO WAY}} \quad (10.32)$$

Los pesos  $w_k$  no tienen una forma fácilmente presentable, pero no existen obstáculos computacionales para obtenerlos. Los pesos y el estimador  $\hat{Y}_{\text{TWO WAY}}$  se calculan fácilmente utilizando software existente como por ejemplo, CALMAR. Este último se describe en [Deville, C. Särndal y Sautory 1993](#).

Los fundamentos teóricos de la calibración para las clasificaciones cruzadas se dan, por ejemplo, en [Deville y C. Särndal 1992](#), donde se hace hincapié en el caso de la respuesta completa. Las extensiones a la falta de respuesta se discuten en [Deville 2000](#) y [Caron, Deville y Sautory 1998](#).

Otro nombre familiar en relación con la información sobre conteos marginales es el método del ratio por ranking, analizado, por ejemplo, en [Oh y Scheuren 1983](#). El algoritmo computacional usado en el método del ratio por ranking proporciona un estimador puntual que no es exactamente idéntico a  $\hat{Y}_{\text{TWO WAY}}$ , como se describe en esta sección, pero generalmente la diferencia numérica entre los dos es intrascendente.

#### 10.4.6 Estimadores simples y su sesgo por falta de respuesta

En los diferentes apartados anteriores de este último capítulo se han visto casos especiales simples del estimador de calibración general  $\hat{Y}_W$ , cuyos pesos  $w_k$  están dados bajo especificaciones estándar por (10.15) y sus generalizaciones. Igualmente, se demuestra que existen evidencias empíricas que relacionan estrechamente el sesgo y la varianza con la información auxiliar incorporada en un estimador. Cuanto más extensa sea la



información auxiliar, menor será el sesgo y la varianza. En este apartado se examinará el sesgo de forma analítica, de la expresión  $\hat{Y}_W$  y obtendremos algunas conclusiones.

El sesgo de  $\hat{Y}_W$ , evaluado conjuntamente con respecto al diseño muestral  $p(s)$  y la distribución de la respuesta  $q(r | s)$ , es

$$\mathbb{B}_{pq}(\hat{Y}_W) = \mathbb{E}_p \left[ \mathbb{E}_q(\hat{Y}_W | s) \right] - Y = \mathbb{E}_{pq}(\hat{Y}_W) - Y$$

El sesgo relativo viene dado por

$$\text{relbias}(\hat{Y}_W) = \frac{\mathbb{B}_{pq}(\hat{Y}_W)}{Y}$$

En el caso más simple tenemos  $\mathbf{x}_k = x_k = 1$  para todo  $k$ , y  $\sum_U x_k = N$ . Entonces  $\hat{Y}_W = \hat{Y}_{\text{EXP}} = N\bar{y}_{r;d}$ , donde  $\bar{y}_{r;d} = \frac{\sum_r d_k y_k}{\sum_r d_k}$ . Se puede obtener fácilmente una aproximación cercana al sesgo reemplazando la suma del numerador y la suma del denominador de  $\bar{y}_{r;d}$  por sus respectivos valores esperados. Teniendo  $E_{pq}(\sum_r d_k y_k) = E_p(\sum_s d_k \theta_k y_k) = \sum_U \theta_k y_k$  y  $E_{pq}(\sum_r d_k) = \sum_U \theta_k$ , por tanto, una aproximación cercana,

$$\mathbb{B}_{pq}(\hat{Y}_{\text{EXP}}) \approx N(\bar{y}_{U;\theta} - \bar{y}_U) \quad (10.33)$$

El sesgo aproximado es por tanto proporcional a la diferencia entre dos medias desconocidas, la media de theta-ponderada  $\bar{y}_{U;\theta} = \frac{\sum_U \theta_k y_k}{\sum_U \theta_k}$  y la media estricta  $\bar{y}_U = \frac{1}{N} \sum_U y_k$ . La theta-ponderada está indicada en  $\bar{y}_{U;\theta}$  (y de forma similar en cantidades ponderadas) por el índice  $\theta$ . El sesgo relativo de  $\hat{Y}_{\text{EXP}}$  es

$$\text{relbias}(\hat{Y}_{\text{EXP}}) \approx \frac{\bar{y}_{U;\theta}}{\bar{y}_U} - 1 \quad (10.34)$$

El lado derecho puede ser positivo o negativo, dependiendo de cuál de las dos medias (desconocidas) sea mayor, la ponderada theta o la estricta. El sesgo relativo es casi cero si las probabilidades de respuesta  $\theta_k$  son constantes en toda la población.

Si usamos también  $\frac{N-1}{N} \approx 1$ , entonces (10.34) conduce a otra expresión reveladora que vincula el sesgo relativo al coeficiente de correlación entre la probabilidad de respuesta y la variable de estudio,

$$\text{relbias}(\hat{Y}_{\text{EXP}}) \approx R_{\theta y U} \text{cv}_{\theta U} \text{cv}_{y U}$$

donde el coeficiente de correlación es

$$R_{\theta y U} = \frac{\text{Cov}_{\theta y U}}{S_{\theta U} S_{y U}} \quad (10.35)$$

y los dos coeficientes de variación son

$$cv_{\theta U} = \frac{S_{\theta U}}{\bar{\theta}_U} \quad (10.36)$$

y

$$cv_{yU} = \frac{S_{yU}}{\bar{y}_U} \quad (10.37)$$

con  $\bar{\theta}_U = \frac{1}{N} \sum_U \theta_k$ ,  $Cov_{\theta y U} = \frac{1}{N-1} \sum_U (\theta_k - \bar{\theta}_U) (y_k - \bar{y}_U)$  y

$$S_{\theta U}^2 = \frac{1}{N-1} \sum_U (\theta_k - \bar{\theta}_U)^2, \quad S_{yU}^2 = \frac{\sum_U (y_k - \bar{y}_U)^2}{N-1}$$

Cuanto mayor sea la correlación  $R_{\theta y U}$ , mayor será el sesgo relativo, en igualdad de condiciones. Incluso una correlación modesta puede resultar un sesgo relativo considerable. Por ejemplo, si  $R_{\theta y U} = 0,4$ ,  $cv_{\theta U} = 0,5$  y  $cv_{yU} = 1,5$ , entonces el sesgo relativo  $(\hat{Y}_{\text{EXP}}) \approx 0,3$  que es inaceptablemente grande.

Para que  $\hat{Y}_{\text{EXP}}$  esté cerca de no tener sesgo, los requerimientos son excepcionalmente fuertes. Se cumplen si las probabilidades de respuesta  $\theta_k$  son constantes en toda la población, algo muy poco probable. La mayoría de los estadísticos que se encargan de realizar encuestas conocen bien esta deficiencia de  $\hat{Y}_{\text{EXP}}$ . Aunque  $\hat{Y}_{\text{EXP}}$  no se recomienda usar, existe cierto interés en calcularlo en una encuesta, aunque sólo sea para ver en qué se diferencia de las mejores estimaciones alternativas.

Un tipo eficaz de información auxiliar es la que permite agrupar los elementos de la población o de la muestra. Idealmente, los grupos deben ser homogéneos con respecto a las probabilidades de respuesta y / o los valores de las variables de estudio. Establecer tal agrupación no es una tarea trivial.

La clasificación unidireccional vista en apartado anterior usa el identificador de grupo  $\gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$  como un vector auxiliar. Para InfoU con  $\mathbf{x}_k = \mathbf{x}_k^* = \gamma_k$ , se obtiene el estimador de ajuste ponderado poblacional  $\hat{Y}_{\text{PWA}}$ , dado por (10.16). Para InfoS con  $\mathbf{x}_k = \mathbf{x}_k^o = \gamma_k$ , se obtiene el estimador de clase ponderado  $\hat{Y}_{\text{WC}}$ , dado por (10.18).

Aproximamos el sesgo mediante un razonamiento similar al utilizado para obtener (10.33) y alternativamente, podemos obtenerlo como un caso especial de la expresión de sesgo mucho más general, tal que:

$$\mathbb{B}_{pq}(\hat{Y}_{\text{PWA}}) \approx \mathbb{B}_{pq}(\hat{Y}_{\text{WC}}) \approx \sum_{p=1}^P N_p (\bar{y}_{U_p; \theta} - \bar{y}_{U_p}) \quad (10.38)$$

donde  $\bar{y}_{U_p;\theta} = \frac{\sum_{U_p} \theta_k y_k}{\sum_{U_p} \theta_k}$  y  $\bar{y}_{U_p} = \frac{1}{N_p} \sum_{U_p} y_k$ . Por tanto  $\hat{Y}_{PWA}$  y  $\hat{Y}_{WC}$  comparten el mismo sesgo aproximado. Es una función de las diferencias entre las medias de los grupos estricta y ponderadas por theta, y es independiente del diseño muestral. Para el sesgo relativo tenemos

$$\text{relbias} \left( \hat{Y}_{PWA} \right) \approx \text{relbias} \left( \hat{Y}_{WC} \right) \approx \frac{\sum_{p=1}^P N_p \bar{y}_{U_p;\theta}}{\sum_{p=1}^P N_p \bar{y}_{U_p}} - 1 \quad (10.39)$$

De forma alternativa, (10.38) puede escribirse como

$$\mathbb{B}_{pq} \left( \hat{Y}_{PWA} \right) \approx \mathbb{B}_{pq} \left( \hat{Y}_{WC} \right) \approx \sum_{p=1}^P K_p \left( \sum_{U_p} \theta_k e_k \right) \quad (10.40)$$

donde  $e_k = y_k - \bar{y}_{U_p}$  para  $k \in U_p$ , y  $K_p = \frac{N_p}{\sum_{U_p} \theta_k}$ . Partiendo de (10.40) seguimos tres condiciones para obtener un sesgo casi nulo:

- 1 Las probabilidades de respuesta  $\theta_k$  son constantes dentro de cada grupo.
- 2 Los residuos satisfacen  $e_k = y_k - \bar{y}_{U_p} = 0$  para todo  $k$  dentro de cada grupo, es decir, el  $y_k$  tiene varianza cero dentro de cada grupo.
- 3 Ni 1 ni 2 se cumplen necesariamente, pero  $\sum_{U_p} \theta_k e_k = 0$  es válido para cada grupo, lo que implica que  $\theta_k$  y  $e_k$  son incorrelados en cada grupo.

Sin embargo, otra forma de expresar el sesgo es

$$\mathbb{B}_{pq} \left( \hat{Y}_{PWA} \right) \approx \mathbb{B}_{pq} \left( \hat{Y}_{WC} \right) \approx \sum_{p=1}^P N_p \bar{y}_{U_p} R_{\theta y U_p} cv_{\theta U_p} cv_{y U_p} \quad (10.41)$$

donde  $R_{\theta y U_p}$ ,  $cv_{\theta U_p}$  y  $cv_{y U_p}$  están definidos por analogía con (10.35)-(10.37), con  $U_p$  reemplazando  $U$ . Por tanto,  $R_{\theta y U_p}$  es la correlación dentro del grupo  $p$  entre la probabilidad de respuesta y la variable  $y$ . Estas correlaciones de grupo juegan un papel clave en la determinación del sesgo. El sesgo también puede ser casi cero si, de manera fortuita, las correlaciones  $R_{\theta y U_p}$  tienen signos alternos para  $p = 1, \dots, P$ , de tal manera que los términos positivos y negativos en (10.41) se cancelarán aproximadamente entre sí.

## Bibliografía

- Bethlehem, J.G. (1988). "Reduction of nonresponse bias through regression estimation". En: *Journal of Official Statistics* 4, págs. 251-260.
- Bethlehem, J.G. y H.M.P. Kersten (1985). "On the treatment of nonresponse in sample surveys". En: *Journal of Official Statistics* 1, págs. 287-300.

- Caron, N., J.C. Deville y O. Sautory (1998). "Estimation de pr ´ ecision de donn ´ ees issues d'enqu ^ etes: document m ´ ethodologique sur le logiciel POULPE". En: *Document de travail de la Direction des Statistiques D ´ emographiques et Sociales* 1.
- Deville, J.C. (2000). "Generalized calibration and application to weighting for non-response". En: *COMPSTAT – Proceedings in Computational Statistics* 1, págs. 65-76.
- Deville, J.C. y C.E. Särndal (1992). "Calibration estimators in survey sampling". En: *Journal of the American Statistical Association* 87, págs. 65-76.
- Deville, J.C., C.E. Särndal y O. Sautory (1993). "Generalized raking procedures in survey sampling". En: *Journal of the American Statistical Association* 88, págs. 1013-1020.
- Kalton, G. y D. Kasprzyk (1986). "The treatment of missing data". En: *Survey Methodology* 12, págs. 1-16.
- Little, R.J.A. (1986). "Survey nonresponse adjustments for estimates of means". En: *International Statistical Review* 54, págs. 139-157.
- Oh, H. L. y F.J. Scheuren (1983). "Weighting adjustment for unit nonresponse. In W. G. Madow, I. Olkin and D. B. Rubin (eds), *Incomplete Data in Sample Surveys*, Vol. 2". En: New York: Academic Press, págs. 167-183.
- Särndal, C.-E. y S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.
- Sweden, Statistics (1980). *Räkna med bortfall (Computation with nonresponse)*. Stockholm: Statistics Sweden.
- Thomsen, I. (1973). "A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data". En: *Statistisk Tidskrift* 11, págs. 278-285.
- (1978). "A second note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data". En: *Statistisk Tidskrift* 16, págs. 191-196.

## Tema 11

**Imputación. ¿Qué es la imputación? Terminología. Múltiples variables de estudio. El enfoque de imputación completa. El enfoque combinado. El enfoque de reponderación completa. Imputación por reglas estadísticas. Imputación por juicio del experto y por datos históricos.**

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

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

D. Haziza (2009). "Imputation and inference in the presence of missing data". En: *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*. Ed. por C.R. Rao y D. Pfeffermann. Amsterdam: North-Holland. Cap. 10, págs. 215-246

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.

### 11.1 Introducción

La falta de respuesta ocurre de forma inevitable en la mayoría, si no en todas, de las encuestas. Esencialmente, los estadísticos distinguen dos tipos de falta de respuesta, la total o falta de respuesta de la unidad, y la parcial o falta de respuesta por ítem. La falta de respuesta total tiene lugar cuando todas las variables de la encuesta son *missing* o no contienen suficiente información que se pueda utilizar.

Por ejemplo, una unidad muestral puede rechazar participar en la encuesta o puede terminar de forma prematura la entrevista. En el último caso, la unidad se identifica como falta de respuesta total incluso si se ha recogido alguna información porque se juzga que es insuficiente. La falta de respuesta parcial tiene lugar cuando *missing* algunos valores (no todos) de las variables de la encuesta. Por ejemplo, una unidad muestral puede rechazar responder a las preguntas sensibles o puede que no conozca la

respuesta de algunas preguntas, o los valores *missing* pueden ser el resultado de que no se verifiquen algunos edits (ver Tema 9).

En el Tema 9 se revisa en profundidad la depuración de datos estadísticos en el proceso estadístico. La falta de respuesta total se trata normalmente con un procedimiento de reponderación. En el procedimiento de reponderación se eliminan las unidades que no han proporcionado información y las ponderaciones de los que sí han respondido<sup>1</sup> al cuestionario se ajustan para compensar estas unidades eliminadas.

La imputación es un proceso en el cual se obtiene un valor artificial para sustituir un valor *missing*. Aunque la imputación a veces se usa para la falta de respuesta total, su uso principal es para compensar la falta de respuesta parcial.

Los principales efectos de la falta de respuesta (total o parcial) incluye: (i) sesgo de los estimadores puntuales, (ii) aumento de la varianza de los estimadores puntuales (ya que el tamaño de la muestra final observada es menor que el tamaño muestral inicialmente planeado), y (iii) sesgo de los estimadores de varianza de los datos completos. El objetivo principal al tratar la falta de respuesta (total o parcial) es reducir el sesgo por falta de respuesta, que tiene lugar si las respuestas de las unidades que responden al cuestionario y las que no responden son distintas con respecto a las variables de la encuesta.

Aunque la imputación múltiple (Rubin 1987) va ganando popularidad en los institutos de estadística, la mayoría de las encuestas usan alguna forma de imputación simple. Por este motivo, en este tema nos centraremos en la imputación simple, que consiste en crear un único valor imputado para sustituir un valor *missing* obteniendo como resultado un único fichero de datos completo. La imputación múltiple consiste básicamente en crear  $M \geq 2$  valores imputados para sustituir un valor *missing*, y como resultado se obtienen  $M$  ficheros de datos completos.

La imputación simple es la más usada a la hora de tratar la falta de respuesta parcial porque presenta las siguientes ventajas: (i) conduce a la creación de un fichero de datos completo, de forma que los resultados de los distintos análisis son consistentes entre sí (en comparación con la imputación múltiple) y (ii) a diferencia de la reponderación para cada ítem, la imputación permite el uso de una única ponderación para todos los ítems (según el diseño muestral original, lo que reduce la complejidad).

Sin embargo, la imputación presenta ciertos riesgos: (i) incluso aunque la imputación conlleva la creación de un fichero de datos completo, las inferencias sólo son válidas si se satisfacen las hipótesis subyacentes sobre el mecanismo de respuesta y/o el modelo de

---

<sup>1</sup>Se traduce *respondent* como el informante que ha respondido completamente al cuestionario, el que no presenta falta de respuesta, en contraste con el informante, que se utiliza como equivalente a unidad muestral que recibe el cuestionario (sin tener en cuenta si responderá o no).

imputación; (ii) algunos métodos de imputación tienden a distorsionar la distribución de las variables de interés (es decir, las variables que son imputadas); (iii) tratar los valores imputados como si fuesen los observados puede implicar una considerable subestimación de la varianza del estimador, especialmente si la tasa de falta de respuesta parcial es considerable; (iv) la imputación de cada ítem de forma separada tiene el efecto de distorsionar las relaciones entre las variables e introducir sesgos incontrolados en las estimaciones.

En ausencia de falta de respuesta, el personal encargado del muestreo normalmente intenta evitar usar procedimientos de estimación cuya validación dependa de la validez de un modelo dado. Para evitar hipótesis sobre la distribución de los datos, las propiedades de los estimadores se basan normalmente en el diseño muestral. Este enfoque es el llamado enfoque basado en el diseño muestral o enfoque de aleatorización sobre el muestreo. Esto no significa que los modelos no sean útiles bajo el enfoque basado en el diseño. De hecho, juegan un papel muy importante en la determinación de un muestreo eficiente y de procedimientos de estimación. El uso de modelos no se puede evitar en presencia de falta de respuesta y las propiedades de los estimadores (puntual y de varianza) (por ejemplo, el sesgo y la varianza) dependerán de la validez de los modelos asumidos.

En consecuencia, la imputación es esencialmente un ejercicio de modelización. La calidad de las estimaciones, por tanto, dependerá de la disponibilidad (en la fase de imputación) de buena información auxiliar y de su uso acertado en la construcción de valores imputados y/o grupos de imputación. La información auxiliar juega un papel muy importante en la encuesta porque permite a los estadísticos usar un muestreo más eficiente y procesos de estimación. También se puede usar para reducir los errores ajenos al muestreo como los errores de falta de respuesta, los errores de cobertura y los errores de medida.

Distinguiremos tres conjuntos de variables auxiliares. El primero es el conjunto de variables de diseño que asumimos que están disponibles para todas las unidades en la población en la fase de diseño. Las variables de diseño se usan habitualmente para estratificar la población o usar alguna forma de muestreo proporcional al tamaño. El segundo conjunto de variables auxiliares se usa para construir valores imputados y/o grupos de imputación y normalmente está relacionado con la variable a imputar y/o la probabilidad de respuesta de esta variable. En otras palabras, estas variables serán útiles para reducir el sesgo por falta de respuesta y posiblemente reducir la varianza por falta de respuesta. Finalmente, asumimos que, en la fase de estimación, se dispone de un conjunto de variables auxiliares (que a menudo se llaman variables de calibrado o *benchmark variables*) para todas las unidades muestrales y para el total poblacional para cada variable que son conocidas en este conjunto. Las variables de calibrado normalmente se especifican para asegurar la consistencia con totales conocidos. Cabe señalar que estos tres conjuntos de variables auxiliares no tienen que ser necesariamente disjuntos, por lo que una variable auxiliar dada se puede usar en distintas fase del proceso de producción.

## 11.2 ¿Qué es la imputación?

### Definición 25

La *imputación* es el procedimiento por el cual los valores *missing* de una o más variables de estudio son 'completados' con sustitutos.

Estos sustitutos se pueden construir de muchas formas, como veremos más adelante en este tema. Los valores imputados son, por definición, artificiales y contienen errores. El *error de imputación* es similar al error de medida (cuando un informante proporciona un valor erróneo en un apartado o en un cuestionario) en el que el valor grabado no es el 'valor real'. Pero a diferencia de los errores de medida, los errores de imputación tienen lugar 'por construcción', ya que los valores que se asignan pueden ser más o menos erróneos.

Otro tipo de construcción de valores artificiales, que usan algunas oficinas de estadística, es la *imputación masiva*. En este procedimiento, los valores son imputados no sólo para los elementos muestrales, sino también para todos los elementos no observados en la población. En este tema no se verá la imputación masiva.

Los valores imputados se pueden clasificar en las siguientes 3 categorías:

1. valores contruidos con la ayuda de una regla estadística de predicción;
2. valores observados de elementos muestrales (similares) que responden (llamados *donantes*);
3. valores contruidos por expertos.

Las categorías 1 y 2 se pueden llamar *reglas estadísticas*, porque usan técnicas estadísticas comunes para producir valores sustitutos. La categoría 1 a menudo se basa en relaciones entre variables, como en la predicción por regresión. Los métodos de la categoría 2 se pueden describir como *basados en donantes*, ya que el valor imputado es 'prestado' por un elemento observado considerado como 'similar' o 'muy similar', basándonos en buenos motivos estadísticos. Los métodos de la categoría 3 se basan en las habilidades y conocimiento del experto sobre el elemento particular que hay que imputar.

A menudo se utiliza otra distinción. Los valores imputados se pueden clasificar como *determinísticos* (cuando si repetimos el procedimiento de imputación los valores imputados toman exactamente el mismo valor) o *aleatorios* (cuando al repetir el procedimiento se obtienen distintos valores imputados). La imputación por regresión es un ejemplo de una regla determinística, mientras que un ejemplo de una regla aleatoria es imputar usando el valor de un elemento observado seleccionado aleatoriamente. Este procedimiento a menudo se llama 'imputación *hot deck*'.



Muchos estadísticos, incluyendo metodólogos y expertos en la materia, observan la imputación con un ojo crítico. Va en contra del sentido común estadístico calcular estadísticas fiables con la ayuda de valores que se sabe desde un principio que pueden ser más o menos correctas. Por otro lado, las razones prácticas para la imputación son muchas veces convincentes.

No hay una evidencia grande de que una imputación cuidadosa sea más perjudicial para las estimaciones que otras fases del proceso de producción estadística. Como en el caso de las ponderaciones, se espera que la imputación dé lugar a estimaciones con, principalmente, un sesgo pequeño y, en segundo lugar, una varianza pequeña.

La imputación puede ser preferible a la reponderación en algunos casos. Un ejemplo surge cuando la población es muy asimétrica, como es el caso de muchas encuestas económicas. El juicio del experto puede entonces ser una forma preferible de obtener un valor imputado 'razonablemente cercano' para una unidad grande e influyente que no contesta. La reponderación puede ser menos interesante para este elemento. El cálculo de ponderaciones siempre implica alguna forma de suavizado, por referencia a 'elementos similares'. Esto no es lo ideal para una unidad muy grande y singular.

En general, una buena imputación requiere un cuidado y una técnica profesional considerable. Los valores imputados deben de ser sustitutos cercanos para los valores no observados. Es una buena política revisar y mejorar de forma continua los métodos de imputación usados en una encuesta.

A menudo se hace una distinción entre *valor imputado* y *valor derivado*. Un valor imputado es un valor construido, insertado en un fichero de datos en lugar de uno que es *missing*. Se obtiene de manera distinta a la observación del elemento a través del cuestionario y no es probable que sea exacto o 'verdadero'. Un valor derivado, por otro lado, también se puede insertar en un fichero de datos, pero se deriva mediante una operación aritmética para todos los elementos de una muestra o de la población. Los valores derivados normalmente se consideran 'valores verdaderos'.

### 11.3 Terminología

Supongamos que de una población  $U$  de tamaño  $N$  extraemos una muestra  $s$  de tamaño  $n$  y, tras recoger la información, tenemos las respuestas completas de un subconjunto  $r$  de la muestra de tamaño  $m$  tal y como se muestra en la Figura 11.1.

Recordemos que existen dos tipos de falta de respuesta: la falta de respuesta total y la falta de respuesta parcial. Se define la falta de respuesta total como aquella en la que el informante no ha contestado a ninguna pregunta del cuestionario. La falta de respuesta parcial se da cuando el informante no ha proporcionado, por lo menos, la respuesta a una pregunta del cuestionario. El conjunto de elementos con una respuesta grabada en al menos uno de los ítems del cuestionario se llamará el conjunto de respuestas. Una

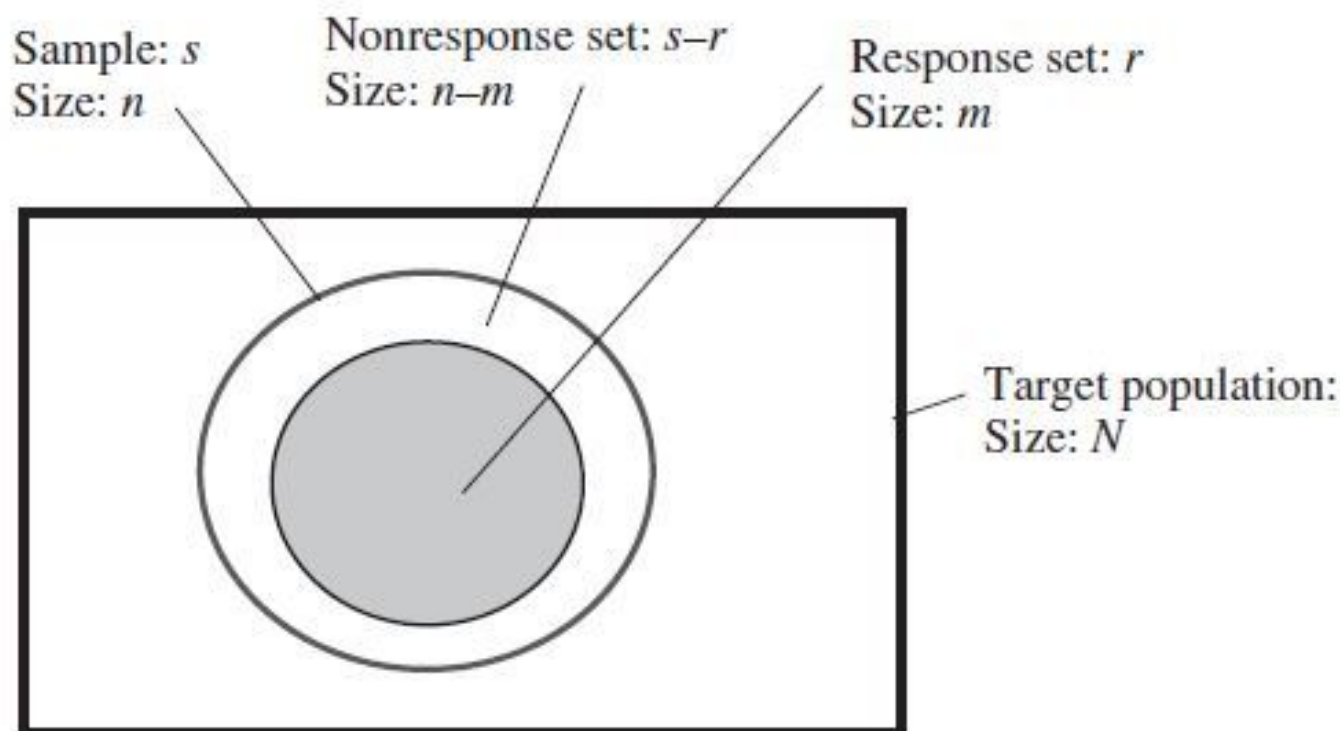


Figura 11.1: Representación de la muestra seleccionada y el conjunto de respuesta, visto como subconjuntos de la población objetivo, que se asume que es idéntica al marco poblacional.

encuesta grande normalmente implica muchas variables de estudio, que se pueden ver afectadas en diversas medidas por la falta de respuesta parcial. Además, generalmente también nos encontraremos con falta de respuesta total.

Los métodos de los que disponemos para tratar la falta de respuesta son la reponderación y la imputación. De forma teórica, esto implica que hay tres posibilidades: utilizar el último, el primero o una combinación de ambos. Por motivos prácticos se puede mostrar preferencia por alguno de estos 3 tipos.

La *reponderación completa* es el enfoque en el cual la ponderación de cada variable de estudio se calcula cada vez. No hay imputación. Este enfoque requiere que se calcule un conjunto de ponderaciones para cada variable de estudio en la encuesta. Supone una ventaja innegable si todas las variables se pueden tratar de forma uniforme en la producción de estadísticas, de forma que se aplique una ponderación común a todas las variables de estudio. La imputación no permite esto.

Una razón que se menciona a menudo en relación a la imputación es que da lugar a un *conjunto de datos rectangular*. La conveniencia de esto es particularmente evidente en una

encuesta grande que involucre muchas variables dependientes<sup>2</sup>  $y$ . Cada registro (cada elemento) define una fila en una matriz de datos con  $I$  columnas, donde  $I$  es el número de variables dependientes  $y$ . Antes de la imputación, la matriz de datos contiene un número de 'agujeros' causados por valores *missing* de  $y$ . Después de la imputación, cada registro tiene  $I$  valores grabados, algunos observados, algunos imputados. El conjunto de datos se ha vuelto rectangular.

Hay dos enfoques que se usan con frecuencia para imputar, en ambos casos se obtienen matrices de datos rectangulares. Son la imputación completa y la combinación de reponderación e imputación.

La *imputación completa* significa que la imputación se usa para tratar tanto la falta de respuesta total como la parcial. Es decir, se imputan todos los valores *missing* para todos los elementos que tengan uno o más valores *missing* de  $y$ . La matriz rectangular de datos completa resultante tiene dimensión  $n \times I$ , donde  $n$  es el tamaño muestral. No hay ajuste de las ponderaciones por falta de respuesta.

*Combinación de reponderación e imputación*, o el *enfoque combinado*, implica que la imputación se restringe a la falta de respuesta parcial. Se imputa para los  $m$  elementos que tienen al menos un valor *missing* de  $y$  pero no todos. La matriz rectangular de datos resultante tiene dimensión  $m \times I$ . La reponderación se aplica entonces para compensar por la falta de respuesta total.

Si es necesario, ambos enfoques también harán un uso apropiado de los pesos de muestreo o de los pesos que incorporan información auxiliar disponible, como en el enfoque de calibrado.

#### *Ejemplo 45. Algunas cuestiones que surgen con la imputación*

El siguiente ejemplo sencillo ilustra parte de la terminología. Supongamos que la encuesta tiene una única variable dependiente  $y$ , de tipo continua, y que cada valor *missing* se imputa por la media de las respuestas de los valores de  $y$ . El conjunto de datos después de la imputación, llamada conjunto de datos completo, consistirá entonces en  $m$  valores realmente observados,  $y_k$ , para  $k \in r$ , y  $n - m$  valores imputados, todos ellos iguales a  $\bar{y}_r = \frac{1}{m} \sum_{k \in r} y_k$ , donde  $r$  es el conjunto de tamaño  $m$  de informantes que han respondido.

Como queda intuitivamente claro, el método no es eficiente y no es recomendable en una encuesta con altas exigencias de calidad. Alcanza el objetivo de obtener un conjunto de datos completo. Pero es fácil identificar varios inconvenientes de este conjunto de

---

<sup>2</sup>En este tema se usará el término variable dependiente para denominar a las variables objetivo incluidas en el cuestionario, puesto que para la imputación desempeñarán el papel de variables dependientes de modelos estadísticos que se precisan predecir para obtener los valores imputados.

datos. Normalmente, ni la tendencia central ni la varianza de estos datos coincidirán con lo que se espera de este conjunto de datos con respuesta completa. La varianza será anormalmente pequeña, porque  $n - m$  valores *missing* se han imputado con el mismo valor, la media de los que han respondido  $\bar{y}_r$ . Además, la tendencia central de estos datos a menudo no reflejarán la tendencia central verdadera de la variable dependiente  $y$ . Si, por ejemplo, unidades con valores de  $y$  grandes responden menos a menudo que las unidades con valores de  $y$  pequeños, entonces la media para el conjunto de datos completo probablemente tendrá una media más pequeña que el conjunto de datos con respuesta completa.

Por contra, si imputamos usando el valor de una unidad que ha respondido seleccionada de forma aleatoria (la 'unidad donante'), entonces la variabilidad del conjunto de datos completo empezará a parecer más 'normal'. Pero la tendencia central tendrá las mismas limitaciones que en el caso de la imputación por la media de los que han respondido. Además, cuando el donante se selecciona de forma aleatoria, corremos el riesgo de imputar el valor de una unidad muy grande para una unidad pequeña. La propiedad deseada de 'cercanía a la realidad' del valor imputado se puede ver gravemente comprometida por este procedimiento simplista. Las estimaciones resultantes corren el riesgo de estar muy sesgadas. ■

## 11.4 Múltiples variables de estudio

Al igual que antes, vamos a denotar por  $U$  y  $s$ , respectivamente, a la población objetivo y a la muestra obtenida a partir de  $U$ . Los pesos muestrales son  $d_k = \frac{1}{\pi_k}$ . Necesitamos notación adicional para incorporar las características de una encuesta que utilice muchas variables de estudio, que pueden verse afectadas de muy distinta forma por la falta de respuesta parcial.

Consideremos una encuesta con  $I$  variables de estudio (o ítems). Podemos centrarnos en una de estas variables, por ejemplo la  $i$ -ésima, que denotaremos por  $y$ , cuyo valor para el elemento  $k$  es  $y_k$ . (Para las otras  $I - 1$  variables de estudio no necesitamos una notación específica). El objetivo es estimar  $Y = \sum_U y_k$ . Si  $y_k$  se observa para cada elemento  $k \in s$ , diremos que la *respuesta es completa* para la variable  $y$ . Sin embargo, en general, la variable  $y$  se verá afectada tanto por la falta de respuesta total como por la parcial.

En línea con la terminología usada en la Sección 11.3,  $r$  denota el conjunto de respuesta para la encuesta. Este subconjunto de la muestra  $s$  consiste en todos los elementos que han respondido a la encuesta. Este subconjunto de la muestra  $s$  está compuesto por todos los elementos que han respondido a una o más de las  $I$  variables de estudio. Un elemento  $k$  que no ha respondido a ninguna de las  $I$  variables pertenece al conjunto  $s - r$  de unidades que no ha respondido. El conjunto de elementos que ha respondido a la variable de estudio  $y$  se denota por  $r_i$  y se llama conjunto de respuesta del ítem  $i$ . Es decir, el valor  $y_k$  se observa para  $k \in r_i$ , donde  $r_i \subseteq r \subseteq s$ . Si  $y_k$  es *missing* y es tratado mediante imputación, denotaremos el valor imputado por  $\hat{y}_k$ . La notación es general;  $\hat{y}_k$  denota un valor que se puede haber obtenido mediante cualquiera de los

métodos estándares. En una encuesta se puede usar más de un método de imputación. Es decir, no todos los valores imputados  $\hat{y}_k$  tienen por qué ser el resultado del mismo método. Cuando  $r_i = r$ , no hay falta de respuesta parcial para la variable  $y$ . Si  $r = s$ , la encuesta no tiene falta de respuesta total, pero  $y$  y el resto de  $I-1$  variables pueden verse afectadas por falta de respuesta parcial. Entonces  $r_i - r$  puede no ser vacío para cada  $i$ .

## 11.5 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 esta sección y en la Sección 11.6, discutiremos la imputación dentro de un marco lo suficientemente general como para cubrir los distintos métodos de imputación que se más se usan.

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:

$$y_{\bullet k} = \begin{cases} y_k, & \text{si } k \in r_i ; \\ \hat{y}_k, & \text{si } k \in s - r_i. \end{cases} \quad (11.1)$$

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 podría haber sido construido por cualquiera de los métodos de imputación que mencionaremos más adelante en el tema.

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  $\bar{y}_{\bullet k} = \frac{1}{n} \sum_{k \in s} y_{\bullet k}$ . Por el contrario, la media que habría sido calculada en el caso de respuesta completa sería  $\bar{y}_s = \frac{1}{n} \sum_{k \in s} y_k$ . 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  $y$ , esto es,  $Y = \sum_{k \in U} y_k$ . 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 (11.1). El resultado se denomina *estimador imputado*. El sistema de ponderación del estimador de respuesta completa se usa sin modificaciones. El conjunto de datos completado (11.1) simplemente reemplaza el conjunto de datos deseado (pero no disponible) que consiste en observaciones reales.

Si el estimador de Horwitz–Thompson,  $\hat{Y}_U^{\text{HT}} = \sum_{k \in s} d_k y_k$ , se utiliza como el estimador de respuesta completa, entonces el equivalente imputado es

$$\hat{Y}_U^{\text{HT}} = \sum_{k \in s} d_k y_{\bullet k} = \sum_{k \in r_i} d_k y_k + \sum_{k \in s-r_i} d_k \hat{y}_k. \quad (11.2)$$

Si el estimador de respuesta completa es el estimador GREG  $\hat{Y}_U^{\text{GREG}} = \sum_{k \in s} d_k g_k y_k$ , los pesos son  $d_k g_k$  para  $k \in s$ . El *input*<sup>3</sup> es  $\sum_{k \in U} \mathbf{x}_k^*$ . El equivalente imputado es

$$\hat{Y}_U^{\text{IGREG}} = \sum_{k \in s} d_k g_k y_{\bullet k} = \sum_{k \in r_i} d_k g_k y_k + \sum_{k \in s-r_i} d_k g_k \hat{y}_k. \quad (11.3)$$

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.

## 11.6 El enfoque combinado

El enfoque combinado es uno de los más utilizados. Recurre a la imputación para la falta de respuesta parcial y a 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 sola etapa<sup>4</sup> para  $k \in r$ :

$$w_k = d_k v_k, v_k = 1 + \left( \mathbf{X} - \sum_{k \in r} d_k \mathbf{x}_k \right)^t \left( \sum_{k \in r} d_k \mathbf{z}_k \mathbf{x}_k^t \right)^{-1} \mathbf{z}_k. \quad (11.4)$$

<sup>3</sup>Se denota por  $\mathbf{x}^*$  al vector auxiliar, y su valor para el elemento  $k$  se denota por  $\mathbf{x}_k^* = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$ , un vector columna con  $J \geq 1$  componentes, construido a partir de una o más variables auxiliares.

<sup>4</sup>Del inglés *single-step calibrated weights*.

Los pesos  $d_k v_k$  tienen incorporada una compensación para la falta de respuesta parcial y la propiedad de calibrado deseada  $\sum_{k \in r} d_k v_k \mathbf{x}_k = \mathbf{X}$ . 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  $Y_U = \sum_{k \in U} y_k$  es entonces

$$\hat{Y}^W = \sum_{k \in r} d_k v_k y_k. \quad (11.5)$$

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 (11.1), también imputamos para  $k \in s - r$ ).

El conjunto completo de datos para la variable  $y$  es  $\{y_{\bullet k} : k \in r\}$ , donde

$$y_{\bullet k} = \begin{cases} y_k, & \text{si } k \in r_i; \\ \hat{y}_k, & \text{si } k \in r - r_i, \end{cases} \quad (11.6)$$

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  $Y = \sum_{k \in U} y_k$  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<sup>5</sup>:

$$\hat{Y}^{IW} = \sum_{k \in r} d_k v_k y_{\bullet k} = \sum_{k \in r_i} d_k v_k y_k + \sum_{k \in r - r_i} d_k v_k \hat{y}_k \quad (11.7)$$

donde los pesos calibrados son  $d_k v_k$ , con  $v_k$  definido en (11.4).

## 11.7 El enfoque de reponderación completa

Una alternativa al enfoque combinado consiste en establecer una dependencia completa en los pesos. En este caso 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

$$w_k = d_k v_{ik}, v_{ik} = 1 + \left( \mathbf{X} - \sum_{k \in r_i} d_k \mathbf{x}_k \right)^t \left( \sum_{k \in r_i} d_k \mathbf{z}_k \mathbf{x}_k \right)^{-1} \mathbf{z}_k. \quad (11.8)$$

<sup>5</sup>Del inglés *item imputed calibration estimator*.



Si  $r_i = r$  en (11.8), usamos la fórmula (11.4) 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 respuesta parcial  $r_i$  a la muestra  $s$ , saltándose  $r$ . Verifican la propiedad de calibrado deseada  $\sum_{k \in r_i} d_k v_{ik} \mathbf{x}_k = \mathbf{X}$ .

El estimador de reponderación completa calibrado<sup>6</sup> para el total  $Y_U = \sum_{k \in U} y_k$  es entonces

$$\hat{Y}^{\text{FW}} = \sum_{k \in r} d_k v_{ik} y_k, \quad (11.9)$$

con  $v_{ik}$  definido en (11.8). 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  $\hat{Y}^{\text{IW}}$  dado por (11.7) y el estimador de reponderación completa  $\hat{Y}^{\text{FW}}$  dado por (11.9) 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 (11.9) son menos en número, numéricamente mayores en media y crean un aumento mayor que los pesos  $d_k v_k$  en (11.7). Un punto importante es que (11.9) y (11.7) son idénticos para algunos tipos de imputación.

**Comentario 40.** 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  $B$ ). Supongamos entonces el caso de tener un atributo definido por una categoría particular de  $A$  y otro de  $B$  y que 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  $Y_U = \sum_{k \in U} y_k$ , el número de elementos de la población con ese atributo.

En el procedimiento totalmente ponderado con pesos definidos por (11.8), 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 (11.8). 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 y, por este motivo, a menudo

<sup>6</sup>Del inglés *fully weighted calibration estimator*.



se prefiere recurrir al enfoque combinado. A menudo este interés por disponer de un único conjunto de pesos se denomina la propiedad multipropósito de los pesos de muestreo (Särndal 2007). ■

## 11.8 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.

### 11.8.1 Imputación por regresión

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

$$\hat{y}_k = \mathbf{x}'_k \hat{\beta}_i, \quad (11.10)$$

donde

$$\hat{\beta}_i = \left( \sum_{k \in r_i} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in r_i} a_k \mathbf{x}_k y_k. \quad (11.11)$$

El vector de coeficientes de regresión  $\hat{\beta}_i$  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  $B_{r_i;a} = \frac{\sum_{r_i} a_k (x_k - \bar{x}_{r_i;a})(y_k - \bar{y}_{r_i;a})}{\sum_{r_i} a_k (x_k - \bar{x}_{r_i;a})^2}$ .

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 (11.10) 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 (11.10) 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  $\bar{y}_{r_i}$ .

### 11.8.2 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  $\ell$  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  $\mathbf{x}_\ell - \mathbf{x}_k$ .

### 11.8.3 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  $x$ . 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.

#### 11.8.4 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.

### 11.8.5 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 (11.10) 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 41.** 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 42.** 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.

## 11.9 Imputación por juicio del experto y por datos históricos

En los institutos de estadística, la imputación a menudo está motivada por querer proporcionar el 'mejor valor imputado posible' basado en el análisis elemento a elemento. La búsqueda es para datos de alta calidad a nivel de microdato, más que a nivel agregado. Para la mayoría de los elementos, en particular para las unidades de tamaño pequeño o mediano, las reglas estadísticas pueden servir bien porque suelen tener características similares. Y pueden ser realizadas por un programa informático. Pero la situación es muy diferente cuando un único elemento puede ser considerado de forma separada para la imputación, sin ningún grupo de referencia aparente o 'elementos similares'. El elemento puede ser muy grande o ser único de alguna otra forma. La imputación especial, por juicio del experto o por datos históricos, se convierte en una necesidad. Esta imputación se reserva en general para una conjunto pequeño de elementos muy influyentes, y es llevado a cabo por expertos, prestando especial atención a

características especiales de cada elemento.

*Ejemplo 46. Imputación estadística en contraste con la imputación especial*

Normalmente, las unidades grandes son influyentes. Su impacto en las estimaciones publicadas puede ser considerable.

Consideremos una encuesta económica en la cual una unidad (empresa/establecimiento) no proporciona la información y tiene que ser imputada. La podemos comparar con el resto de unidades en su mismo grupo de la clasificación de actividades económicas (p.ej. CNAE <sup>7</sup>), pero una simple imputación basada en la media de las respuestas para este grupo daría lugar a un error de imputación negativo excesivo,  $\hat{y}_k - y_k$ , para esta unidad grande. De forma similar, la media de los que han respondido para un grupo de unidades definidas como 'grandes' en algún sentido general (cifra de negocios, empleados, etc.) también puede ser erróneo, porque una unidad que es 'grande' en una actividad económica puede ser muy diferente a otra que es 'grande' en otra actividad económica.

La media de los que responden de un grupo puede ser una imputación poco satisfactoria. El valor imputado de un donante identificado como el 'vecino más cercano' también puede resultar poco adecuado, porque en la cola derecha de la distribución incluso el elemento más cercano puede ser numéricamente muy diferente.

Un mejor enfoque es a menudo una combinación de los datos históricos y el juicio, subjetivo, del experto. Uno puede empezar examinando la serie de los últimos datos proporcionados por la unidad, especialmente los más recientes, y ajustarlos en función del mejor juicio sobre tendencias en la actividad económica y en la economía en general. La justificación es que las unidades grandes a menudo son tan únicas que ninguna de las reglas estadísticas es probable que 'se acerquen'. Pero incluso contando con el mejor juicio y con los mayores habilidades tenemos que considerar la posibilidad de que un error notable en una estimación del total se pueda atribuir a un error excesivo de imputación en una unidad única pero muy influyente.

Además, la imputación por juicio del experto introduce el problema (no tratado en este temario) de la estimación de la varianza en presencia de valores imputados. ■

---

<sup>7</sup>Clasificación Nacional de Actividades Económicas [https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614](https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614).

## Bibliografía

- Haziza, D. (2009). "Imputation and inference in the presence of missing data". En: *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*. Ed. por C.R. Rao y D. Pfeffermann. Amsterdam: North-Holland. Cap. 10, págs. 215-246.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
- Särndal, C.-E. (2007). "The calibration approach in survey theory and practice". En: *Survey Methodology* 33, págs. 99-119.
- Särndal, C.-E. y S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.



## Tema 12

**Control del secreto estadístico. Conceptos y definiciones: Control del secreto estadístico, datos tabulares, microdatos, riesgo y utilidad. Un enfoque al control del secreto estadístico: por qué la protección de la confidencialidad es importante, características clave y usos de los datos, riesgos contra los que la protección es necesaria, métodos de control del secreto, implementación.**

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

C. Skinner (2009). "Statistical disclosure control for survey data". En: págs. 381-396 en (Pfeffermann y Rao 2009)

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.

### 12.1 Conceptos y definiciones: Control del secreto estadístico, datos tabulares, microdatos, riesgo y utilidad

#### El problema del control del secreto estadístico

El secreto estadístico es uno de los principios generales de la Función Estadística Pública (LFEP) <sup>1</sup> y el Capítulo III de la LFEP se titula 'Del secreto estadístico'. En esta ley se recoge qué tipo de datos están amparados por el secreto estadístico y cuáles no, quién tiene la obligación de preservar el secreto estadístico así como las sanciones por el incumplimiento del deber del secreto estadístico.

De acuerdo con la LFEP, a los informantes de encuestas se les garantiza que sus respuestas serán tratadas de forma confidencial. Estas garantías se refieren a la forma en que sus respuestas serán gestionadas en el organismo que realice la encuesta y también a la naturaleza del resultado estadístico de la encuesta. La confidencialidad estadística

---

<sup>1</sup><https://www.boe.es/buscar/doc.php?id=BOE-A-1989-10767>

también forma parte del Código de Buenas Prácticas de las Estadísticas Europeas ([Eurostat 2017](#)), que incluye el Principio 5 relacionado con la '**Confidencialidad Estadística**' según el cual: 'La privacidad de los informantes (hogares, empresas, administraciones y otros encuestados), la confidencialidad de la información que proporcionan y su uso exclusivo con fines estadísticos están totalmente garantizados'.

En este tema se analizarán los métodos que aseguren que no se producirán estadísticas que permitan la identificación de un informante. En el contexto de este tema, el *Control del Secreto Estadístico* (CSE, en inglés *Statistical Disclosure Control* SDC) se refiere a la metodología usada, durante el diseño de los resultados estadísticos de la encuesta, para proteger la confidencialidad de las respuestas de los informantes. El secreto estadístico engloba también la seguridad informática y los protocolos para la gestión de los datos dentro de la oficina de estadística que no se verán en este tema.

Para la difusión de operaciones estadísticas se pueden usar varios tipos de *resultados estadísticos*. El más tradicional son las tablas de resultados, como totales, medias y proporciones. La publicación de estas estimaciones a partir de encuestas a hogares e individuos no se ha considerado como un gran problema de confidencialidad, ya que la protección viene proporcionada por el muestreo. Sin embargo, las tabulaciones de resultados de los distintos tipos de encuestas económicas (a empresas o establecimientos), sí puede conllevar algún riesgo, especialmente por tratamiento de la confidencialidad de información en el caso de grandes empresas en celdas de tablas en las cuales la fracción de muestreo es del 100 %.

Aunque la forma tradicional de publicación de todas las estimaciones de una operación estadística en un listado cerrado de variables sigue satisfaciendo muchas necesidades en los últimos años, ha habido una demanda creciente de resultados más flexibles, donde el conjunto de parámetros poblacionales de interés no esté preestablecido. Hay varios motivos por lo que esto puede no ser posible. El análisis de datos es un proceso iterativo, y lo más interesante que pueden ofrecer los análisis puede que sólo quede claro después de un análisis exploratorio inicial de los datos. Teniendo en cuenta el gasto que implica realizar encuestas, es normal que se busque que el mayor número de usuarios tengan acceso al mayor número de datos. Pero normalmente es imposible especificar de antemano todos los posibles usuarios. Una forma natural de proporcionar resultados flexibles es hacer disponibles los microdatos, de forma que los usuarios puedan llevar a cabo los análisis estadísticos que les interesen.

Sin embargo, la publicación de tales microdatos plantea serias cuestiones de protección de la confidencialidad. Los análisis estadísticos de datos no precisan el conocimiento de la identificación de las unidades muestrales, lo que permite la creación de ficheros *anonimizados* de microdatos eliminando de los mismos los nombres, direcciones e información de contacto de individuos o empresas. El problema, sin embargo, puede surgir si la anonimización básica no es suficiente para asegurar la confidencialidad.

## Los conceptos de confidencialidad, secreto y riesgo de identificación

Para ser más precisos sobre el significado de 'protección de confidencialidad', empezaremos el tema con algunas definiciones. Hablaremos de tres partes implicadas: (1) el *informante* que proporcionan los datos, (2) el *INE* que recoge los datos, publica resultados estadísticos y diseña la estrategia de CSE, y (3) el hipotético *intruso* que tiene acceso a estos resultados y busca usarlos para obtener información sobre los informantes.

Un concepto importante es la *identificación*, que tiene lugar si el intruso consigue relacionar un individuo conocido (u otra unidad) a un registro individual de microdatos u otro elemento de los resultados estadísticos. Otro concepto es la *identificación de atributos*, que tiene lugar si el intruso puede determinar el valor de alguna variable del cuestionario para un individuo identificado (u otra unidad) usando los resultados estadísticos.

De forma general, el *identificación predictiva* ocurriría si el intruso pudiese predecir el valor de alguna variable del cuestionario para un individuo identificado con alguna incertidumbre. Al evaluar el potencial para la identificación a partir de un determinado resultado estadístico, normalmente nos referimos al *riesgo de identificación*, que se define como la probabilidad de identificación respecto a determinadas causas de incertidumbre. O el término se puede usar de manera general para enfatizar no sólo la incertidumbre acerca de la identificación potencial sino también el daño que puede surgir en la identificación ([Lambert 1993](#)). La *confidencialidad* de las respuestas proporcionadas por un informante se puede decir que están protegidas si el riesgo de identificación para este informante y sus respuestas es suficientemente bajo. (Para una mayor discusión sobre definiciones véase [Duncan y Lambert 1986](#); [Duncan y Lambert 1989](#); [Skinner 1992](#)).

## Aproximaciones a la protección de la confidencialidad

Si se considera que el riesgo de identificación no es suficientemente bajo, entonces será necesario usar algún método para reducirlo. Para conseguirlo existen dos aproximaciones, que se denominan *entorno seguro* y *datos seguros* ([Marsh, Dale y Skinner 1994](#)). La aproximación por entorno seguro impone restricciones sobre el conjunto de posibles usuarios de los resultados estadísticos y/o en las formas en que los resultados pueden ser usados. Por ejemplo, los usuarios puede que sólo tengan acceso a los microdatos en un entorno seguro. La aproximación por datos seguros, por otro lado, implica algunas modificaciones sobre el resultado estadístico. Por ejemplo, el grado de desagregación geográfica en un fichero de microdatos en una encuesta social puede estar limitada a un áreas que no contenga menos de una cierta cantidad de unidades en muestra. En este tema nos centraremos en la aproximación por datos seguros y de forma general nos referiremos a los métodos que modifican los resultados estadísticos como métodos de CSE.

## Métodos de CSE, utilidad y calidad de los datos

Los métodos de CSE varían dependiendo de la forma del resultado estadístico. Algunas aproximaciones son las siguientes:

- *Reducción de detalles*, por ejemplo, el número de categorías de una variable categórica se puede reducir en una tabla de clasificación cruzada (o tabla de contingencia) o en los microdatos.
- *Supresión*, por ejemplo, el resultado en una tabla puede ser sustituido por un asterisco, indicando que el dato ha sido sustituido por motivos de confidencialidad.

En cada uno de estos casos, el método de CSE dará lugar a una *pérdida de información* para el usuario de los resultados. Por tanto, el método reducirá el número de parámetros poblacionales para los cuales un usuario puede obtener estimaciones. Otro tipo de métodos de CSE pueden no modificar el número de parámetros que pueden ser estimados, sino afectar a la *calidad* de las estimaciones que se obtienen. Por ejemplo, si se añade *ruido aleatorio* a la variable cifra de negocios para proteger la confidencialidad, entonces se puede introducir sesgo o un aumento de la varianza en las estimaciones asociadas. El término general *utilidad* se puede usar para incluir tanto la información proporcionada por los resultados, por ejemplo, el rango de las estimaciones o los análisis que se pueden producir, como la calidad de esta información, por ejemplo, los errores de las estimaciones. Habría que tener en cuenta, por supuesto, que los datos procedentes de encuestas están sujetas a muchas fuentes de errores, incluso antes de la aplicación de métodos de CSE, y que, por tanto, el impacto de estos métodos en la calidad de los datos tiene que ser considerada en este contexto.

Generalmente, hay que considerar la utilidad desde la perspectiva de un *usuario* de los resultados estadísticos, que representa la cuarta parte a añadir a las tres partes a las que ya nos referimos anteriormente: los informantes, la oficina de estadística, y el intruso.

### CSE como un problema de optimización: el equilibrio entre el riesgo y la utilidad

El punto clave en el CSE es cómo tratar el equilibrio entre el riesgo de identificación y la utilidad. En general, cuanto más se reduzca el riesgo de identificación con un método de CSE, menor será la utilidad esperada del resultado. Este equilibrio puede formularse como un problema de optimización. Sea  $D$  el conjunto de datos (anonimizados) de la encuesta y sea  $f(D)$  el resultado, obtenido del uso del método de CSE. Sea  $R[f(D)]$  una medida del riesgo de identificación del resultado, y sea  $U[f(D)]$  una medida de la utilidad del resultado. Entonces, el problema básico del CSE se puede representar como el siguiente problema de optimización con restricciones:

dados  $D$  y  $\varepsilon$ , encontrar un método de CSE,  $f(\cdot)$ , tal que  
se maximice  $U[f(D)]$ , sujeto a  $R[f(D)] < \varepsilon$

donde:

$f(\cdot)$  es el *método de CSE*. A continuación veremos algunos de estos métodos;

$R(.)$  es la *función del riesgo de identificación*. Veremos distintas formas en que se puede definir esta función;

$U(.)$  es la *función de utilidad*. Tampoco será fácil de definir como una función escalar, dados los múltiples usos de los resultados;

$\varepsilon$  es el *máximo riesgo aceptable*. En principio, se puede esperar que el INE proporcione este valor como garantía para los informantes. Sin embargo, en la práctica, los INEs encuentran muy difícil especificar un valor de  $\varepsilon$  distinto de cero, es decir, sin riesgo de identificación.

Teniendo en cuenta todas estas dificultades a la hora de especificar  $R(.)$  y  $U(.)$  como funciones escalares y en especificar un valor para  $\varepsilon$ , este problema de optimización se utiliza principalmente como motivación conceptual. En la práctica, los distintos métodos de CSE se pueden evaluar y comparar considerando los valores de medidas alternativas de riesgo y utilidad. Para medidas dadas de cada, a veces puede ser útil construir un mapa RU (Duncan, Fienberg y col. 2001), donde una medida de riesgo se dibuja frente a una medida de utilidad para un conjunto de métodos CSE. Se espera que los puntos del mapa dibujen una relación positiva entre riesgo e utilidad, pero se puede ver que, para una valor de riesgo dado, algunos métodos tienen una mayor utilidad que otros, y por tanto serán preferibles. Este enfoque evita tener que asumir un único valor de  $\varepsilon$ .

## **12.2 Un enfoque el control del secreto estadístico: por qué la protección de la confidencialidad es importante, características clave y usos de los datos, riesgos contra los que la protección es necesaria, métodos de control del secreto, implementación**

### **12.2.1 Aplicación a resultados en tablas**

#### **Riesgo de identificación en las encuestas sociales**

Los principales desarrollos en métodos de CSE para resultados tabulados han sido motivados por el riesgo de identificación cuando se usa el 100 % de la muestra, como es el caso de los censos o de los datos administrativos. Las tablas de frecuencia basadas en estas fuentes de datos incluyen a menudo totales pequeños, tales como cero o uno, por ejemplo, en tablas de número de defunciones por áreas por causa de muerte. Estas tablas pueden dar lugar a la identificación de la identidad, por ejemplo, si es de conocimiento público que alguien a muerto, entonces puede ser posible identificar a esa persona si el total es uno en una tabla de defunciones usando alguna característica conocida de esa persona. También puede ocurrir la identificación de atributos. Por ejemplo, puede ser posible obtener la causa de muerte de una persona si en la tabla de contingencia se cruza esta causa con otras variables potenciales conocidas por el intruso.

En encuestas sociales, sin embargo, el uso de muestreo reduce considerablemente el riesgo de este tipo de identificación por dos razones. La primera, la presencia de muestras requiere diferente tipos de resultados. Por tanto, los registros de las tablas para variables categóricas tienden a ser porcentajes elevados (posiblemente dentro de dominios

definidos por filas o columnas) en lugar de totales sin elevar. Incluso si un usuario de la tabla pudiese calcular los totales de la celda (por ejemplo, porque se ha proporcionado la muestra base y la encuesta usa los mismos pesos), las oficinas de estadística además se asegurará de que las celdas publicadas no contienen totales muy pequeños, en los que las estimaciones puedan ser consideradas poco fiables por los errores de muestreo. Por ejemplo, la oficina suprimirá celdas en las que el total muestral esté por debajo de un determinado umbral, por ejemplo, 50 personas en una encuesta social. Esto debería de evitar el tipo de situaciones que más preocupan con el 100 % de los datos. A veces las oficinas usan técnicas de estimación en pequeñas áreas <sup>2</sup> en dominios con totales muestrales pequeños y estas técnicas también pueden servir para reducir el riesgo de identificación.

En segundo lugar, la presencia de muestreo reduce la precisión con la que un intruso puede conseguir identificación predictiva. Por ejemplo, supongamos que un intruso puede obtener a partir de una tabla que, entre 100 informantes que caen dentro de un determinado dominio, 99 de ellos tienen un determinado atributo y supongamos que el intruso conoce a alguien de la población que cae dentro de este dominio. Entonces el intruso no puede predecir que esta persona tiene el atributo con probabilidad 0,99, ya que esta persona necesita no ser un informante y la predicción está sujeta a la incertidumbre muestral. Esta conclusión depende, sin embargo, de que las identidades de los informantes se mantengan confidenciales por parte de la oficina, previniendo que el intruso sepa si la persona conocida es un informante. Esto fue denominado como *conocimiento de respuesta* por (Bethlehem, Keller y Pannekoek 1990). En general es muy importante que las oficinas adopten esta práctica ya que mantiene la confidencialidad de las identidades de los informantes sin afectar a la utilidad estadística de los resultados. En algunos casos excepcionales, puede ser difícil conseguir esto de forma completa.

Por estos motivos, la pérdida de confidencialidad no será, en general, un motivo de preocupación en la publicación de tablas de estimaciones en encuestas sociales en las que las probabilidades de inclusión de la muestra sean pequeñas, por ejemplo inferiores a 0.1. (Véase también [Statistical Methodology 2005](#))

### **Riesgo de identificación en las encuestas económicas**

Una forma común de publicar los resultados en una encuesta económica consiste en una tabla de clasificación cruzada de totales estimados, con características de los establecimientos. Cada estimación será de la forma  $\hat{Y}_c = \sum_s w_i I_{ci} y_i$ , donde  $w_i$  es el peso muestral,  $I_{ci}$  es una variable indicadora para la celda  $c$  en la clasificación cruzada e  $y_i$  es el valor de la variable para el establecimiento  $i$ -ésimo en la muestra  $s$ . Por ejemplo,  $y_i$  puede ser una medida de una variable (cifra de negocios, número de trabajadores) y las celdas se pueden obtener por el cruce con la clasificación con la actividad económica y una medida de tamaño.

---

<sup>2</sup>Véase el Tema 8. *Estimación en dominios. Los métodos básicos de estimación en dominios. Condicionamiento sobre el tamaño muestral del dominio. Dominios pequeños: estimadores sintéticos.* en el bloque *Producción Estadística Oficial: Métodos Avanzados*



Es tales circunstancias, la definición de confidencialidad será en forma de riesgo de identificación. El riesgo de identificación para una celda específica  $c$  se debe definir bajo las siguientes condiciones e hipótesis:

- el intruso es uno de los establecimientos/empresas con datos en la celda y tiene por objetivo predecir el valor  $y_i$  para otro establecimiento  $o$ , en general, el intruso consiste en un *conjunto* de  $m$  de los  $N_c$  establecimientos con datos en la celda con el mismo objetivo predictivo;
- el intruso conoce las identidades de todos los establecimientos dentro de la celda (ya que, por ejemplo, pueden representar a su competencia).

Bajo estas hipótesis, la identificación predictiva se puede decir que ocurre si el intruso es capaz de predecir el valor de  $y_i$  con un grado específico de precisión. Para clarificar la notación de precisión, nos centramos en el caso en que las unidades en una celda caen todas dentro de un estrato exhaustivo. Por tanto,  $w_i = 1$  cuando  $I_{ci} = 1$ , de forma que  $\hat{Y}_c = \sum_{U_c} y_i$ , donde  $U_c$  es el conjunto de establecimientos en la celda  $c$  y  $N_c$  es el tamaño de  $U_c$ . En este caso, el intruso no se enfrenta con incertidumbre debida al muestreo y esto se tiene que considerar como el peor caso posible.

#### • Identificación predictiva en ausencia de muestreo

En este caso la predicción se considera normalmente desde una perspectiva determinista y se representa por un intervalo (entre un límite superior y uno inferior) dentro del cual el intruso sabe que se encuentra el valor  $y_i$ . La precisión de la predicción se representa por la diferencia entre el valor verdadero y el del intervalo. Se supone que si el intruso obtiene la predicción combinando información previa con el valor publicado  $\hat{Y}_c$ .

Una aproximación para especificar la información previa se usa en la *regla prior-posterior* (Willenborg y Waal 2001), también llamada la regla  $pq$ , que depende de dos constantes,  $p$  y  $q$ , fijadas por la oficina. La constante  $q$  se usa para especificar la precisión de la predicción basada únicamente en la información previa. Bajo la regla  $pq$  se asume que el intruso puede deducir el valor  $y_i$  para cada establecimiento en la celda con un  $q$  %. Así, la oficina asume que, antes de la publicación de los resultados, el intruso puede saber que un valor  $y_i$  cae dentro del intervalo  $[(\frac{1-q}{100})y_i, \frac{1+q}{100})y_i]$ . La combinación de esta información previa con el resultado  $\hat{Y}_c = \sum_{U_c} y_i$  puede ser utilizada por el intruso para obtener mejores límites para el valor verdadero. Por ejemplo, sean los estadísticos ordenados  $y_{(1)} < y_{(2)} < \dots < y_{(N_c)}$  y supongamos que el intruso es el establecimiento con el segundo mayor valor,  $y_{(N_c)-1}$ . Entonces, el intruso puede establecer una cota superior para el mayor valor  $y_{(N_c)}$  restando su propio valor junto con la suma de las cotas inferiores para  $y_{(1)}, \dots, y_{(N_c-2)}$  de  $\hat{Y}_c$ . La precisión de la predicción usando este límite superior viene dada por la diferencia entre esta cota superior y el verdadero valor  $y_{(N_c)}$ , que es  $(\frac{q}{100}) \sum_{i=1}^{N_c-2} y_i$ . Esta celda se dirá que *sensible* bajo la regla  $pq$ , es decir, se considera como *reveladora*, si esta diferencia es inferior que el  $p$  % del verdadero valor, esto es, si

$$\left(\frac{p}{100}\right)y_{(N_c)} - \left(\frac{q}{100}\right)\sum_{i=1}^{N_c-2} y_i > 0. \quad (12.1)$$

La expresión en la parte izquierda de la ecuación (1) es un caso especial de una *medida lineal de sensibilidad*, que en general será de la forma  $R_c = \sum_{i=1}^{N_c} a_i y_{(i)}$ , donde  $a_i$  son pesos específicos. La celda se dice sensible si  $R_c > 0$ . En este caso, se considerará que tendrá lugar la identificación predictiva. Un caso especial de la regla  $pq$  es la *regla del  $p\%$* , que se da cuando  $q = 100$ , es decir, se asume que no hay información previa. Otra medida de sensibilidad lineal comúnmente usada es la  $(n, k)$  o *regla de dominancia*. (Véase [Willenborg y Waal 2001](#); [Statistical Methodology 2005](#)).

- **Identificación predictiva en presencia de muestreo**

En un caso más general, el estrato no es exhaustivo. En este caso,  $\hat{Y}_c$  estará sujeto a un error muestral y, en general, habrá una protección a la identificación adicional, ya que el intruso desconoce si un establecimiento (que no pertenezca al grupo de establecimientos que quieren obtener la información) está o no en la muestra. La definición de riesgo en estas condiciones necesita más investigación. Se presentan algunas ideas en la sección 6.2.5 del ([Willenborg y Waal 2001](#)).

Un enfoque alternativo estocástico basado en modelos debe asumir que antes de la publicación de los resultados la información previa sobre  $y_i$  puede representarse por un modelo de regresión lineal dependiente de los valores de las covariables disponibles públicamente  $x_i$  con una varianza residual específica. La distribución predictiva de  $y_i$  dado  $x_i$  podría entonces actualizarse usando el(los) valor(es) de  $y_i$  conocidos por el intruso y el valor publicado  $\hat{Y}_c$ , que se asumirá que sigue una distribución  $\hat{Y}_c \sim N(Y_c, v(\hat{Y}_c))$ , donde  $v(\hat{Y}_c)$  es la estimación de la varianza de  $\hat{Y}_c$ . La identificación predictiva se puede entonces medir en función de la varianza residual resultante en la predicción de  $y_i$ .

## Métodos de CSE para resultados tabulados

Si una celda en una tabla se considera sensible, es decir, el valor de la celda representa un riesgo inadmisiblemente alto de identificación (pérdida de confidencialidad), se puede usar alguno de los métodos de CSE.

- **Redefinición de celdas**

Las celdas se redefinen para eliminar las celdas sensibles, por ejemplo, combinando celdas sensibles con otras o combinando categorías de las variables cruzadas. Esto también se denomina *rediseño de tablas* ([Willenborg y Waal 2001](#)).

- **Supresión de celdas**

El valor de una celda sensible se elimina. Dependiendo de la naturaleza de la tabla y de marginales publicadas, también puede ser necesario eliminar los valores de celdas 'complementarias' para prevenir que un intruso sea capaz de deducir el



valor de la celda a partir de otros valores en la tabla. Hay mucha literatura sobre los distintos enfoques acerca de la elección de las celdas complementarias que aseguran la protección. (Véase por ejemplo, [Willenborg y Waal 2001](#); [Cox 2001](#); [Giessing 2001](#), y las referencias que figuran en ellos).

- **Modificación de celdas**

Los valores de la celda se pueden modificar de alguna forma. Normalmente será necesario modificar no sólo los valores de las celdas sensible sino también valores en algunas celdas no sensible complementarias, por la misma razón que en la supresión de celdas. La modificación puede ser determinística (por ejemplo, [Cox 2001](#)), o estocástica (por ejemplo la Sección 9.2 de [Willenborg y Waal 2001](#)).

Un método sencillo es el *redondeo*, donde los valores de las celdas modificadas se multiplican por un entero de referencia dado (capítulo 9 de [Willenborg y Waal 2001](#)). Este método se utiliza en tablas de frecuencias obtenidas del 100 % de los datos, pero también se puede usar a tablas de totales estimados, donde el entero de referencia se puede elegir de acuerdo con las magnitudes de los totales estimados. En lugar de reemplazar los valores de la celda por valores seguros, también es posible sustituirlos por intervalos, definidos por las cotas inferior y superior ([Giessing y Dittrich 2006](#)).

El método del *ajuste tabular controlado* ([Cox, Kelly y Patil 2004](#)) determina los valores de las celdas modificadas entre unos límites tales que la tabla permanezca aditiva y se verifican determinadas propiedades estadísticas y de seguridad.

- **Modificación de microdatos previa a la tabulación**

En lugar de modificar los valores de las celdas, los microdatos con los que se obtienen los resultados pueden ser perturbados, por ejemplo, añadiendo ruido, y luego obtener la tabla con los valores perturbados ([Evans, Zayatz y Slanta 1998](#)).

Los resultados de una encuesta incluirán habitualmente muchas tablas. Aunque los métodos ya mencionados se pueden aplicar de manera separada a cada tabla, estas aproximaciones no tienen en cuenta los posibles riesgos de identificación adicionales consecuencia de la combinación de información de distintas tablas, en particular de marginales comunes. Para protegerse de estos riesgos adicionales se plantean nuevas consideraciones para el CSE. Además, el conjunto de tablas que constituyen los resultados estadísticos no es necesariamente fijo. Con los desarrollos de la difusión online, hay una demanda creciente de creación de tablas que puedan responder de una forma más flexible a las necesidades de los usuarios. Esto implica la necesidad de considerar métodos de CSE que no sólo proteja cada tabla de manera separada, como se ha visto en las secciones anteriores, sino también contra el riesgo que surge de posibles sucesiones alternativas de tablas publicadas (véase por ejemplo [Dobra, Karr y Sanil 2003](#)).

## 12.2.2 Aplicación a microdatos

### Evaluación del riesgo de identificación

Supongamos que la oficina está considerando la publicación del fichero anonimizado de microdatos, donde los registros del fichero se corresponden con las unidades básicas de análisis y cada registro contiene el listado de variables de la encuesta. Este fichero también puede incluir identificadores para unidades de análisis de alto nivel, por ejemplo, identificadores de hogares donde la unidad básica son los individuos, así como la información necesaria para el análisis de la encuesta tales como los pesos de muestreo e identificadores de las unidades primarias de muestreo (PSU del inglés *primary sampling unit*).

Supongamos que un intruso puede enlazar un registro en el fichero con alguna fuente de datos externa de unidades conocidas usando algunas variables, que están incluidas tanto en el fichero de microdatos como en la fuente externa. Estas variables se denominan a menudo como *variables clave* o variables identificadoras. Hay varias formas de definir el riesgo de identificación en estas condiciones. (Véase, por ejemplo, [Paass 1988](#); [Duncan y Lambert 1989](#)). Un enfoque común, a menudo motivado por la naturaleza de la garantía de confidencialidad, es considerar una forma de *riesgo de identificación* ([Bethlehem, Keller y Pannekoek 1990](#)), considerando la posibilidad de que el intruso sea capaz de asociar los microdatos con una unidad conocida. Esta definición de riesgo será únicamente apropiada si los registros en los microdatos se pueden asociar de forma coherente con las unidades de la población. Cuando los microdatos están sujetos a alguna forma de CSE, esto no se puede verificar (por ejemplo, si los registros publicados se obtienen combinando los registros originales), y en este caso puede ser más apropiado considerar alguna definición de identificación predictiva.

Existen varias estrategias para evaluar el riesgo de identificación, pero todas dependen de las hipótesis sobre la naturaleza de las variables clave. un enfoque es realizar un experimento empírico, cruzando los microdatos propuestos con otra fuente de datos, que se considera como un sustituto para la fuente de datos que tenga el intruso. Una vez hechas las hipótesis sobre las variables clave, la oficina de estadística puede usar métodos de record linkage<sup>3</sup> para enlazar unidades entre los dos conjuntos de datos. El riesgo entonces se puede medir en términos del número de unidades para las cuales tienen lugar las correspondencias junto con una medida de la calidad de coincidencia (en términos de proporciones de falsos positivos y negativos). Este experimento, por tanto, precisa que la oficina tenga información que permita establecer de forma precisa qué unidades son comunes a las dos fuentes y cuáles no.

El punto clave en este enfoque es cómo construir un conjunto de datos sustituto realista que el intruso pueda tener, que pueda tener una parte común con las unidades incluidas en los microdatos y conocer la naturaleza de ese solapamiento. En algunos casos

---

<sup>3</sup>Tema 14

puede haber una fuente de datos disponible. [Blien, Wirth y Muller 1992](#) proporciona un ejemplo de una fuente de datos que contiene un listado de personas en determinadas profesiones. Otra posibilidad puede ser otra encuesta realizada por la oficina de estadística, aunque a menudo las oficinas controlan las muestras para impedir este solapamiento. Incluso si hay solapamiento, por ejemplo con un censo, determina con precisión qué unidades son comunes y cuáles no puede consumir muchos recursos. Por tanto, el uso rutinario de un enfoque es difícil.

En ausencia de otro conjunto de datos, el INE puede considerar un experimento de reidentificación, en el cual el fichero de microdatos se combina consigo mismo de forma parecida, después de la aplicación de algún método CSE. Este enfoque tiene la ventaja de que no depende del modelo, pero es posible que el riesgo de reidentificación sea sobreestimado si los efectos de protección del muestreo y del error de medida no son realistas.

En lo que queda de sección, consideraremos un tercer enfoque, que de nuevo sólo necesita datos del fichero de microdatos, pero que realizar hipótesis teóricas, especialmente sobre el modelo, para estimar el riesgo de identificación. Lo mismo que en el experimento de reidentificación, este enfoque debe hacer hipótesis sobre cómo se miden las variables clave en los microdatos y por parte del intruso sobre las unidades conocidas usando la información externa.

Una hipótesis que simplifica, pero que es el 'peor caso', es el de que las variables clave sean grabadas de forma idéntica en los microdatos y en el exterior. Nos referiremos a esto como a la *hipótesis de falta de error de medida*, ya que el error de medida en cada una de las fuentes de datos se puede esperar que invaliden esta hipótesis. Si por lo menos una de las variables clave es continua y se considera la hipótesis de falta de error de medida, entonces un intruso que observe un emparejamiento exacto entre los valores de las variables clave en los microdatos y en las unidades conocidas podría concluir con probabilidad uno que el emparejamiento es correcto. En otras palabras, el riesgo de identificación sería uno. Si por lo menos una de las variables clave es continua y se supone que puede tener lugar el error de medida, entonces el riesgo generalmente será menor que uno. Además, un enfoque de emparejamiento exacto no es obviamente sensato y se debe considerar una clase más amplia de métodos de *record linkage*. (Véase [Fuller 1993](#), para la evaluación del riesgo de identificación bajo algunas hipótesis del modelo de error de medida.)

En la práctica, las variables casi nunca se graban de manera continuada en los microdatos de una encuesta social. Por ejemplo, la edad raramente se grabará como una variable continua, sino como entera. Y a partir de ahora nos restringiremos al caso de variables clave categóricas. Por simplicidad, nos limitamos al caso de correspondencia exacta, aunque se pueden usar métodos de *record linkage* más generales. Nos centraremos en los ficheros de microdatos, en los que los únicos métodos de CSE que se han aplicado son la recodificación de las variables clave o el (sub)muestreo.

## Medidas del riesgo de identificación a nivel del fichero

Consideremos una población  $U$  con  $N$  unidades (que típicamente serán individuos) y supongamos que el fichero de microdatos está formado por los registros de una muestra  $s \subset U$  de tamaño  $n \leq N$ . Asumimos que la posibilidad de identificación ocurre si un intruso consigue acceso a los microdatos e intenta enlazar un registro a nivel de microdato con información externa sobre una unidad conocida usando los valores de  $m$  variables clave categóricas  $X_1, \dots, X_m$ . (Cabe señalar que  $s$  y  $X_1, \dots, X_m$  se definen después de la aplicación de (sub)muestreo o la recodificación, respectivamente, como métodos CSE al fichero de microdatos original.)

Denotemos por  $X$  a las variables  $X_1, \dots, X_m$ , cuyos valores, que se corresponden con una posible combinación de categorías de las variables clave de  $X$ , denotaremos por  $k = 1, \dots, K$ , siendo  $K$  el número de valores clave de  $X$ . Cada uno de estos valores clave se corresponde con una posible combinación de categorías de variables clave. Bajo la hipótesis de falta de error de medida, la identificación de la identidad es una preocupación si un registro es único en la población respecto a las variables clave. Un registro con valor clave  $k$  se dirá que es *único en la población* si  $F_k = 1$ , donde  $F_k$  denota el número de unidades en  $U$  con valor clave  $k$ . Si un intruso observa un emparejamiento con un registro con un valor clave  $k$ , sabe que el registro es único en la población y puede considerar la hipótesis de falta de error de medida, entonces el intruso puede inferir que el emparejamiento es correcto.

Como una medida sencilla del riesgo de identificación, debemos por tanto considerar algún resumen de la extensión de la unicidad de la población. En muestreo, es normal definir parámetros de interés a nivel poblacional y esto nos puede llevar a definir la medida como la proporción poblacional  $N_r/N$ , donde  $N_r = \sum_k I(F_k = r)$ ,  $r = 1, 2, \dots$  es la frecuencia poblacional de frecuencias. Desde la perspectiva del riesgo de identificación, sin embargo, estamos interesados en el riesgo para un fichero de específico de microdatos y es natural que la medida de riesgo dependa de la muestra. Por tanto, supondremos que el riesgo será más alto si la muestra seleccionada contiene una proporción más alta de unidades identificables inusuales que si en la muestra la proporción es menor. Por tanto, una medida más natural a nivel de fichero es la proporción de únicos en la población incluidos en la muestra. Denotemos por  $f_k$  el equivalente muestral de  $F_k$ , entonces esta medida se puede expresar de la siguiente forma:

$$\mathbb{P}(PU) = \sum_k I(f_k = 1, F_k = 1)/n. \quad (12.2)$$

Se puede argumentar que el denominador de esta proporción se puede hacer más pequeño, ya que sólo los registros que son únicos en la población pueden ser únicos en la muestra, debido a que  $f_k \leq F_k$ , es decir, tienen un valor clave  $k$  tal que  $f_k = 1$ . Por tanto, una medida más conservadora sería considerar

$$\mathbb{P}(PU|SU) = \sum_k I(f_k = 1, F_k = 1)/n_1 \quad (12.3)$$

donde  $n_1$  es el número de únicos en la muestra y, de forma más general,  $n_r = \sum_k I(f_k = r)$  es la frecuencia muestral de frecuencias. (Para más información sobre la proporción de valores únicos en la muestra que son únicos en la población véase [Fienberg y Makov 1998](#)).

Se puede argumentar ([Skinner y Elliot 2002](#)) que estas medidas pueden ser excesivamente optimistas, ya que sólo recogen el riesgo que existe para los únicos en la población y no para otros registros con  $F_k \geq 2$ . Si un intruso observa un emparejamiento en un valor clave con frecuencia  $F_k$ , entonces (sujeto a la hipótesis de que no haya error de medida) la probabilidad de un emparejamiento correcto es  $1/F_k$  bajo la hipótesis de intercambiabilidad de que el intruso habrá seleccionado con la misma probabilidad cualquiera de las  $F_k$  unidades de la población. Una medida alternativa de riesgo se obtiene extendiendo esta idea de probabilidad de emparejamiento correcto para los diferentes valores clave. De nuevo, poniéndonos en el peor de los casos, será normal prestar atención a los únicos en la muestra. Una medida surge de suponer que los inicios del intruso con los microdatos, es igualmente probable seleccionar cualquier muestra único de la muestra y luego emparejarlo con la población. La probabilidad de que el emparejamiento resultante sea correcto es la media de  $1/F_k$  entre los únicos en la muestra:

$$\theta_s = \frac{1}{n_1} \sum_k \frac{I(f_k = 1)}{F_k} \quad (12.4)$$

Otra medida es

$$\theta_U = \frac{\sum_k I(f_k = 1)}{\sum_k F_k I(f_k = 1)} \quad (12.5)$$

que es la probabilidad de una correspondencia correcta bajo el supuesto en el que el intruso busca aleatoriamente entre la población y encuentra una correspondencia con un único en la muestra.

Las cuatro medidas anteriores dependen de  $f_k$  y de  $F_k$ . La oficina de estadística que realice la encuesta será capaz de determinar las cantidades muestrales  $f_k$  a partir de los microdatos, pero las cantidades poblacionales  $F_k$  serán en general desconocidas. Por tanto, resulta de interés ser capaz de hacer inferencia sobre las medidas de los datos de la muestra.

[Skinner y Elliot 2002](#) demostraron que, bajo muestreo de Bernoulli con probabilidad de inclusión  $\pi$ , un estimador insesgado de  $\theta_U$  es  $\hat{\theta}_U = \frac{n_1}{[n_1 + 2(\pi^{-1} - 1)n_2]}$ . También proporcionaron un estimador consistente para la varianza asintótica de  $\hat{\theta}_U - \theta_U$ . [Skinner y Carter 2003](#) demostraron que un estimador consistente de  $\theta_U$  para un diseño complejo arbitrario es  $\hat{\theta}_U = \frac{n_1}{[n_1 + 2(\bar{\pi}_2^{-1} - 1)n_2]}$ , donde  $\bar{\pi}_2^{-1}$  es la media de la inversa de las probabilidades de inclusión  $\pi_i^{-1}$  para las unidades  $i$  con valores clave para los cuales  $f_k = 2$ . También proporcionaron un estimador consistente para la varianza asintótica de  $\hat{\theta}_U - \theta_U$  bajo el muestreo de Poisson.

Se pueden hacer hipótesis de diseño más fuertes, en particular asumiendo que los  $F_k$  son independientes e idénticamente distribuidos como:

$$F_k | \lambda_k \sim \mathcal{P}(\lambda_k) \quad (12.6)$$

donde los  $\lambda_k$  son independientes e idénticamente distribuidos, es decir, los  $F_k$  se distribuyen según una Poisson. Una opción manejable para la distribución de  $\lambda_k$  es la distribución gamma (Bethlehem, Keller y Pannekoek 1990) aunque no parece que se ajuste bien en algunas aplicaciones de datos reales. La log-normal proporciona un mejor ajuste (Skinner y Holmes 1993). Samuels 1998 discutió la estimación de  $\mathbb{P}(PU|SU)$  basada en un modelo Poisson-Dirichlet. Una conclusión general parece ser que los resultados pueden ser en cierto modo sensibles a la elección del modelo, especialmente cuando la fracción de muestreo decrece, y que  $\theta_U$  se puede estimar de manera más robusta que las otras tres medidas.

### Medidas del riesgo de identificación a nivel del registro

Un problema con las medidas a nivel de fichero es que los principios de protección de confidencialidad a menudo buscan evitar la identificación de *cualquier* individuo, es decir, requieren que el riesgo esté por debajo de un umbral para cada registro, y estos objetivos pueden no ser abordados de manera adecuada con medidas globales del tipo (2)-(5). Puede resultar más natural considerar medidas a nivel de registro, es decir, medidas que pueden tomar diferentes valores para cada registro de microdatos. Estas medidas pueden ayudar a identificar las partes de la muestra en las que el riesgo es mayor y se necesita mayor protección (Lambert 1993). Aunque las medidas a nivel de registro pueden proporcionar mayor flexibilidad cuando evalúan si formas específicas de resultados son 'transparentes', son más difíciles de estimar que las medidas a nivel de fichero.

Se han propuesto varios enfoques para la estimación de medidas a nivel de medidas a nivel de registro. Para variables clave continuas, Fuller 1993 mostró cómo evaluar la probabilidad de identificación a nivel de registro en presencia de ruido añadido, bajo la hipótesis de normalidad. (Véase también Paass 1988; Duncan y Lambert 1989). Consideremos a continuación métodos para variables categóricas (Skinner y Holmes 1993).

Consideremos un registro de microdatos con valor clave  $X$ . Supongamos que el registro es único en la muestra, es decir, con un valor clave  $k$  para el cual  $f_k = 1$ , ya que este valor es el que se espera que presente más riesgo. Supongamos que el intruso observa una correspondencia exacta entre este registro y una unidad conocida de la población. Consideramos la hipótesis de falta de error de medida de forma que habrá  $F_k$  unidades en la población que potencialmente podrán emparejarse con el registro. También asumimos que no hay conocimiento de respuesta (véase 12.2.1). La probabilidad de que el emparejamiento observado sea correcto es

$$\mathbb{P}(\text{emparejamiento correcto} | \text{emparejamiento exacto}, X = k, F_k) = 1/F_k \quad (12.7)$$



donde la distribución de probabilidad con respecto al diseño está bajo un esquema simétrico de muestreo, como un muestreo aleatorio simple o un muestreo de Bernoulli. (De forma alternativa, podría ser respecto a un mecanismo estocástico usado por el intruso, que selecciona cualquiera de las  $F_k$  unidades con igual probabilidad). Esta probabilidad está condicionada al valor clave  $k$  y a  $F_k$ .

En la práctica, sólo observamos las frecuencias muestrales  $f_k$  y no  $F_k$ , así que reescribimos la ecuación anterior:

$$\mathbb{P}(\text{emparejamiento correcto} | \text{emparejamiento exacto}, X = k, f_k) = \mathbb{E}(1/F_k | k, f_k = 1) \quad (12.8)$$

Esta esperanza se calcula con respecto tanto al esquema de muestreo como al modelo que genera  $F_k$ , como por ejemplo el modelo de Poisson visto en (12.6). Una medida alternativa, que se centra en el riesgo de carácter único poblacional es

$$\mathbb{P}(F_k = 1 | k, f_k = 1). \quad (12.9)$$

Las expresiones (12.8) y (12.9) se pueden generalizar a cualquier registro de los microdatos con  $f_k > 1$ . Una diferencia entre las probabilidades en (12.8) y (12.9) y las de las secciones anteriores es que aquí condicionamos al valor clave del registro  $X = k$ . Por tanto, aunque debemos asumir  $F_k | \lambda_k \sim \mathcal{P}(\lambda_k)$ , como en (12.6), nos gustaría condicionar sobre el valor clave  $k$  en particular cuando consideramos la distribución de  $\lambda_k$ . De otra forma, si  $\lambda_k$  está idénticamente distribuido, como en la sección previa, entonces obtendríamos la misma medida de riesgo para todos los registros (únicos en la muestra). Un modelo natural es el modelo log-lineal:

$$\log(\lambda_k) = z_k \beta, \quad (12.10)$$

donde  $z_k$  es un vector de variables indicadoras que representan los principales efectos y las interacciones entre las variables clave  $X_1, \dots, X_m$ , y  $\beta$  es un vector de parámetros desconocidos.

Se pueden encontrar expresiones para las medidas de riesgo en (12.8) y (12.9) en términos de  $\beta$  en Skinner y Holmes 1993. Son necesarias hipótesis sobre el esquema de muestreo para estimar  $\beta$ . Bajo muestreo de Bernoulli con probabilidad de inclusión  $\pi$ , se sigue a partir de (12.6) que  $f_k | \lambda_k \sim \mathcal{P}(\pi \lambda_k)$ . Asumiendo también (12.10) puede estimarse por métodos de máxima verosimilitud. Una extensión sencilla de este argumento también se aplica bajo muestreo de Poisson en el que las probabilidades de inclusión  $\pi_k$  pueden variar en relación con las variables clave, por ejemplo, si una variable de estratificación se incluye entre las variables clave. En este caso, tenemos  $f_k | \lambda_k \sim \mathcal{P}(\pi_k \lambda_k)$ . Skinner y Shlomo 2008 discutieron métodos para la especificación del modelo en (12.10). Skinner 2007 discutió la posible dependencia de la medida sobre el método de búsqueda usado por el intruso.

## Métodos CSE

Veamos algunos métodos de CSE para microdatos.

- **Transformación de variables para reducir el detalle**

Las variables clave categóricas se pueden transformar combinando categorías. Variables clave continuas se pueden *agrupar* en intervalos para dar lugar a variables categóricas especificando puntos de corte entre los cuales los intervalos definen categorías. Proporcionada la transformación, este método tiene la ventaja de que la reducción de la utilidad está clara para los usuarios de los datos, que pueden sufrir la pérdida de información pero la validez de los análisis no se ve afectada.

- **Perturbación estocástica de variables**

Los valores de las potenciales variables clave se perturban de forma estocástica. En el caso de variables continuas, la perturbación puede implicar la *adicción de ruido* (Fuller 1993). En el caso de variables categóricas, la perturbación puede consistir en la clasificación errónea, denominada *Método de Postaleatorización* (PRAM<sup>4</sup>) por Gouweleeuw y col. 1998. La perturbación puede ser llevada a cabo de forma que se preserven características específicas de los microdatos, por ejemplo, las medias y las desviaciones típicas de las variables en los microdatos perturbados pueden ser los mismos que en los microdatos originales, pero en la práctica habrá inevitablemente características de los microdatos que no se podrán reproducir. Por ejemplo, la correlación estimada entre una variable perturbada y una sin perturbar. En principio, esto podrá permitir análisis válidos aunque habrá una pérdida de precisión y las desventajas prácticas serán significantes.

Una alternativa es proporcionar a los usuarios los detalles precisos del método de perturbación, incluyendo los valores de los parámetros, como la desviación típica del ruido e las entradas en la matriz de clasificación errónea, de forma que se pueda 'deshacer' el impacto de la perturbación cuando se lleven a cabo los análisis. Véase, por ejemplo, Hout y Heijden 2002 en el caso de PRAM o Fuller 1993 en el caso de ruido añadido. En principio, esto puede permitir validar los análisis aunque habrá una pérdida de precisión y las desventajas prácticas son significativas.

- **Microdatos sintéticos**

Este enfoque es similar al previo, excepto que su objetivo es evitar la necesidad de métodos especiales de análisis. En su lugar, los valores de las variables en la fila son sustituidos por valores generados por un modelo de forma que se diseña para el análisis de los datos sintéticos, como si fueran los datos originales, para generar estimaciones puntuales consistentes. El modelo se obtiene ajustándolo a

---

<sup>4</sup>del inglés *Postrandomization Method*



los microdatos originales. Para poder calcular errores válidos así como estimadores puntuales consistentes, [Raghunathan, Reiter y Rubin 2003](#) propusieron que se generasen múltiples copias de los microdatos sintéticos de forma que se pudiera usar la metodología de imputación.

- **Perturbación selectiva**

A menudo el problema se centra únicamente en los registros que se consideran que tienen demasiado riesgo y se espera que la utilidad sea mayor si sólo se perturba un subconjunto de registros en riesgo. Es posible crear valores missing en estos registros, denominado *supresión local* por [Willenborg y Waal 2001](#), o tanto crear valores missing como sustituir por valores imputados, denominado *blanquear e imputar* por [Statistical Methodology 2005](#). Un problema importante con estos métodos es que es probable que creen sesgos si los valores objetivo son inusuales y los usuarios de los datos no serán capaces de cuantificar estos sesgos.

- **Intercambio de registros**

Los métodos previos se centran en la perturbación de valores de las variables para todos o para un conjunto de registros. El método de intercambio de registros, en cambio, implica que los valores de una o más de una variable clave se intercambian entre registros. La elección de los registros entre los cuales los valores se intercambian se puede controlar de forma que ciertas frecuencias bivariantes o multivariantes se mantengan, en particular intercambiando sólo registros que compartan ciertas características ([Willenborg y Waal 2001](#), sección 5.6). En general, sin embargo, no será posible controlar todas las relaciones multivariantes y el intercambio de registros puede perjudicar la utilidad de forma análoga a la clasificación errónea ([Skinner y Shlomo 2008](#)).

- **Microagregación**

Este método es pertinente para variables continuas, como las de encuestas económicas, y en su forma principal consiste en ordenar los valores de cada variable y formar grupos de un tamaño específico  $k$  (el primero grupo contiene los  $k$  menores valores, el segundo grupo los  $k$  siguientes menores valores, y así consecutivamente). El método reemplaza los valores por la media del grupo, de forma separada para cada variable. Una ventaja del método es que la modificación de los datos será mayor para los valores outliers, que también se consideran como los que tienen más riesgo. Sin embargo, es difícil para los usuarios evaluar el sesgo provocado por el método en los análisis.

Los métodos CSE se aplicarán en general después de la fase de depuración de la encuesta, en la que se comprueba que se verifican una serie de edits. La aplicación de algún método de CSE puede, sin embargo, provocar que no se verifiquen estos edits.

### CSE para pesos de muestreo y otra información de diseño

Los pesos de muestreo y otra información de diseño compleja también se suelen publicar junto con los microdatos de la encuesta con el fin de que se puedan llevar a cabo análisis válidos. Sin embargo, es posible que esta información de diseño pueda contribuir a la pérdida de confidencialidad. Por tanto, la variable de diseño de la encuesta (la variable que se usa para estratificar) también puede ser una variable clave. Hay que tener en cuenta que esto no significa que los pesos de muestreo no se puedan publicar; significa que la evaluación del riesgo de identificación deberían de tener en cuenta la información que los pesos de muestreo pueden transmitir. [Willenborg y Waal 2001](#) propusieron algunos enfoques de ajuste de los pesos para reducir el riesgo.

Además de la publicación de los pesos de muestreo, es frecuente publicar el estrato o los identificadores de las UPM o los identificadores duplicados. Un enfoque sencillo para evitar publicar los UPM o identificadores duplicados es proporcionar en su lugar la información sobre los efectos de diseño, las funciones de varianza generalizadas o el uso de pesos replicados mediante un bootstrap ajustado.

## 12.3 Conclusiones

El desarrollo de la metodología de CSE continúa, estimulada por una gran cantidad de problemas prácticos y de innovaciones continuas en las formas en que los datos de las encuestas se usan, sin señales de que las preocupaciones sobre la confidencialidad decrezcan. Estos métodos se benefician de los modelos estadísticos o el record linkage, que también han evolucionado mucho en los últimos años.

## Bibliografía

- Bethlehem, J.G., W.J. Keller y J. Pannekoek (1990). "Disclosure control for microdata". En: *Journal of the American Statistical Association* 85, págs. 38-45.
- Blien, U., H. Wirth y M. Muller (1992). "Disclosure risk for microdata stemming from official statistics". En: *Statistica Neerlandica* 46, págs. 69-82.
- Cox, L.H. (2001). "Disclosure risk for tabular economic data. En: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies". En: Elsevier, págs. 167-183.
- Cox, L.H., J.P. Kelly y R. Patil (2004). "Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (Eds.), Privacy in Statistical Databases. Lecture Notes in Computer Science". En: Elsevier, págs. 87-98.
- Dobra, A., A.F. Karr y A.P. Sanil (2003). "Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues". En: *Statistics and Computing* 13, págs. 363-370.
- Duncan, G.T., S.E. Fienberg, R. Krishnan nad R. Padman y S.F. Roehrig (2001). "Disclosure limitation methods and information loss for tabular data. En: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies". En: Elsevier, págs. 135-166.

- Duncan, G.T. y D. Lambert (1986). "Disclosure-limited data dissemination". En: *Journal of the American Statistical Association* 81, págs. 10-28.
- (1989). "The risk of disclosure for microdata". En: *Journal of Business and Economic Statistics* 7, págs. 207-217.
- Eurostat (2017). *Código de Buenas Prácticas de las Estadísticas Europeas*. URL: [https://ine.es/ine/codigobp/codigo\\_2017.pdf](https://ine.es/ine/codigobp/codigo_2017.pdf).
- Evans, T., L. Zayatz y J. Slanta (1998). "Using noise for disclosure limitation of establishment tabular data". En: *Journal of Official Statistics* 14, págs. 537-551.
- Fienberg, S.E. y U.E. Makov (1998). "Confidentiality, uniqueness and disclosure limitation for categorical data". En: *Journal of Official Statistics* 14, págs. 385-397.
- Fuller, W.A. (1993). "Masking procedures for microdata disclosure limitation". En: *Journal of Official Statistics* 9, págs. 383-406.
- Giessing, S. (2001). "Nonperturbative disclosure control methods for tabular data. En: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*". En: Elsevier, págs. 185-213.
- Giessing, S. y S. Dittrich (2006). "Harmonizing table protection: results of a study. In: Domingo-Ferrer, J., Franconi, L. (Eds.), *Privacy in Statistical Databases. Lecture Notes in Computer Science*". En: Springer, págs. 35-47.
- Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg y P.-P. deWolf (1998). "Post randomisation for statistical disclosure control: theory and implementation". En: *Journal of Official Statistics* 14, págs. 463-478.
- Hout, A. Van den y P.G.M. Van der Heijden (2002). "Randomized response, statistical disclosure control, and misclassification: a review". En: *International Statistical Review* 70, págs. 269-288.
- Lambert, D. (1993). "Measures of disclosure risk and harm". En: *Journal of Official Statistics* 9, págs. 313-331.
- Marsh, C., A. Dale y C. Skinner (1994). "Safe data versus safe setting: access to microdata from the British Census". En: *International Statistical Review* 62, págs. 35-53.
- Paass, G. (1988). "Disclosure risk and disclosure avoidance for microdata". En: *Journal of Business and Economic Statistics* 6, págs. 487-500.
- Pfeffermann, D. y C.R. Rao, eds. (2009). *Sample Surveys: Design, Methods and Applications*. Elsevier.
- Raghunathan, T.E., J.P. Reiter y D.R. Rubin (2003). "Multiple imputation for statistical disclosure limitation". En: *Journal of Official Statistics* 19, págs. 1-16.
- Samuels, S.M. (1998). "A Bayesian species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment". En: *Journal of Official Statistics* 14, págs. 373-383.
- Skinner, C. (1992). "On identification disclosure and prediction disclosure for microdata". En: *Statistica Neerlandica* 46, págs. 21-32.
- (2007). "The probability of identification: applying ideas from forensic science to disclosure risk assessment". En: *Journal of the Royal Statistical Society, Series A* 170, págs. 195-212.
- (2009). "Statistical disclosure control for survey data". En: págs. 381-396.

- Skinner, C. y R.G. Carter (2003). "Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling". En: *Survey Methodology* 29, págs. 177-180.
- Skinner, C. y M.J. Elliot (2002). "A measure of disclosure risk for microdata". En: *Journal of the Royal Statistical Society, Series B* 64, págs. 855-867.
- Skinner, C. y D.J. Holmes (1993). "Modelling population uniqueness". En: *International Seminar on Statistical Confidentiality Proceedings, European Community Statistical Office*, págs. 175-199.
- Skinner, C. y N. Shlomo (2008). "Assessing identification risk in survey microdata using log-linear models". En: *Journal of American Statistical Association* 103, págs. 989-1001.
- Statistical Methodology, Federal Committee on (2005). "Report on Statistical Disclosure Limitation Methodology". En: *Statistical Policy Working Paper 22, 2nd version, U.S. Office of Management and Budget, Washington, D.C.*
- Willenborg, L. y T. de Waal (2001). *Elements of Statistical Disclosure Control*. New York: Springer.

## Tema 13

**Difusión de datos: Presentación de estadísticas. Introducción. Transmitir el mensaje. Visualización de las estadísticas. Tablas. Gráficos. Mapas. Técnicas de visualización emergentes. Cuestiones de accesibilidad.**

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

UNECE (2009). *Cómo hacer comprensibles los datos. Parte 2: Una guía para presentar estadísticas*. URL: [https://unece.org/DAM/stats/documents/writing/MDM\\_Part2\\_Spanish.pdf](https://unece.org/DAM/stats/documents/writing/MDM_Part2_Spanish.pdf)

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.

### 13.1 Introducción

Una imagen vale más que mil palabras. Los patrones de comportamiento de los datos suelen ser más evidentes cuando estos se representan gráficamente. Hay muchas formas de presentar los datos, desde gráficos de barras simples hasta diagramas de dispersión, mapas temáticos y gráficos animados más complejos.

Una publicación efectiva de los datos requiere de una combinación de texto, tablas y gráficos para maximizar la solidez en la comunicación de los diversos tipos de información. En este tema se verá cuáles son las mejores prácticas para acercar las estadísticas a la realidad de personas no relacionadas con la estadística.

### 13.2 Transmitir el mensaje

Las notas de prensa son, a menudo, el vehículo a través del cual la organización estadística comunica las principales conclusiones de sus programas estadísticos y analíticos a la audiencia objetivo, que es probablemente el público general. El texto es el vehículo

principal para explicar las conclusiones, destacando las tendencias y proporcionando información contextual.

Éstos son algunos consejos de utilizad para empezar a trabajar en un texto.

1. **Audiencia objetivo: primera decisión.** Lo que el público quiere es lo que se le debería ofrecer. Hoy en día, los organismos de estadística han desarrollado un número significativo de lectores directos a través de sus sitios web, correo electrónico y otras formas de distribución basadas en Internet. Esto significa que se están comunicando con varios tipos de audiencia simultáneamente: el público general, los usuarios de datos, analistas de datos, profesores universitarios, estudiantes y así sucesivamente, cada uno con sus propias necesidades de información. Si lo que se desea es atender a diversas audiencias, se debe seleccionar el método más adecuado para llegar a cada una de ellas, transmitiendo el mensaje a través de canales adecuados y utilizando técnicas de comunicación apropiadas.
2. **Entender el contexto en el que se comunica.** Es importante entender el contexto en el que se está comunicando. Se deben utilizar las herramientas del lenguaje, la estructura y la presentación más apropiadas para transmitir el mensaje. Las siguientes secciones ilustran cómo hacerlo.
3. **Primero y sobre todo, encontrar el relato.** Para que los datos sean comprensibles, es importante encontrarle significado a los números. Es vital que las organizaciones de estadística permanezcan imparciales y garanticen la confidencialidad de los encuestados y de pequeñas subpoblaciones. El texto debe situar los resultados más importantes y significativos en el contexto de las tendencias tanto a corto como a largo plazo.
4. **Escribir en estilo periodístico.** Utilizar el estilo de escritura adoptado por los periodistas: la 'pirámide invertida'. Consiste en presentar los hechos más importantes primero, seguido por los puntos secundarios en orden decreciente de importancia.  
Evitar comenzar el texto con la metodología y terminarlo con una conclusión. Se pueden poner los elementos clave sobre la metodología en una nota a los lectores. La conclusión debería ir en el encabezamiento o párrafo inicial.  
El encabezamiento es el elemento más importante del texto. Debe contar algo acerca de los datos y resumir la trama de forma concisa, clara y sencilla, situando la historia en su contexto.  
No cargar al lector con demasiados números en el cuerpo del texto y utilizar exclusivamente los datos esenciales expresados de forma redondeada. Los números menos importantes deben ser relegados a cuadros adjuntos.
5. **Prestar atención a la estructura.** Estructurar el texto de manera que cada componente tenga sentido por sí mismo y contribuya también a la historia general que se desea contar. Los subtítulos son una herramienta eficaz para reforzar la organización de un comunicado, dividiéndolo en secciones manejables y significativas.
6. **Lenguaje: escribir claro, conciso y sencillo.** Un lenguaje sencillo es la clave de una comunicación exitosa. Esto no se aplica sólo al texto, también es relevante

para las tablas, imágenes y gráficos.

Algunos consejos para una escritura clara: Utilizar frases cortas; Tratar una idea en cada frase; Dividir las frases largas; Iniciar cada párrafo con lo principal del mensaje; Usar párrafos cortos; Mantener una escritura concisa.

7. **Evaluar el impacto: análisis de los medios de comunicación.** Se puede trabajar directamente con los medios de comunicación para garantizar que los mensajes informen con precisión e imparcialidad. Los medios de comunicación son la manera más sencilla, barata y efectiva de ofrecer el mensaje estadístico a una amplia audiencia.

### 13.3 Visualización de las estadísticas

#### Por qué una imagen vale más que mil palabras

Todos hemos oído el viejo dicho: 'una imagen vale más que mil palabras'. Como ya se ha comentado, una de las mejores técnicas para hacer comprensibles los datos es la representación de los números mediante imágenes. Esto puede hacer mucho más fácil apreciar un patrón o exponer ciertos patrones que de otro modo podrían quedar ocultos.

Se pueden mostrar los datos de muchas maneras diferentes, desde sencillos gráficos de barras a diagramas de dispersión más complejos, mapas temáticos y pirámides de población animadas.

#### La visualización como parte esencial del desarrollo de trabajos estadísticos

La presentación efectiva de los datos debe ser una parte esencial del proceso de producción de estadísticas. La visualización está incluida en la fase de difusión del Generic Statistical Business Process Model (Tema ??).

Es mucho más fácil entender las estadísticas presentadas mediante un gráfico o mapa, que en largas listas de números, siempre y cuando las representaciones gráficas están correctamente realizadas.

Las representaciones deben ilustrar las tendencias y las relaciones de forma rápida y sencilla. Pero hay que tener cuidado, porque una mala representación de la información estadística puede resultar engañosa.

#### La influencia histórica de un escocés en la visualización de datos

William Playfair (1759-1823), fue el primero en producir representaciones gráficas de datos estadísticos en formas que ahora a todos nos resultan familiares.

Es importante evitar que las representaciones sean muy llamativas y distorsionen el contenido. Hay que recordar que la tecnología es sólo una herramienta. No se deben



agregar notas inútiles ni elementos retorcidos sólo porque se tenga la capacidad de hacerlo, sino que hay que mantener la sencillez del mensaje para el lector.

### Cuestiones básicas sobre la percepción humana

Nuestra capacidad de hacer observaciones visuales rápida y fácilmente, se basa en la capacidad del cerebro para percibir regularidades e irregularidades. Gran parte de esta capacidad funciona inconscientemente ya que la comparación se produce casi antes de empezar a pensar en ello.

El mensaje para los estadísticos es el siguiente: se debe tener cuidado al elaborar representaciones visuales de observaciones estadísticas, pues el contexto en que los resultados se presentan pueden distorsionar la percepción del usuario.

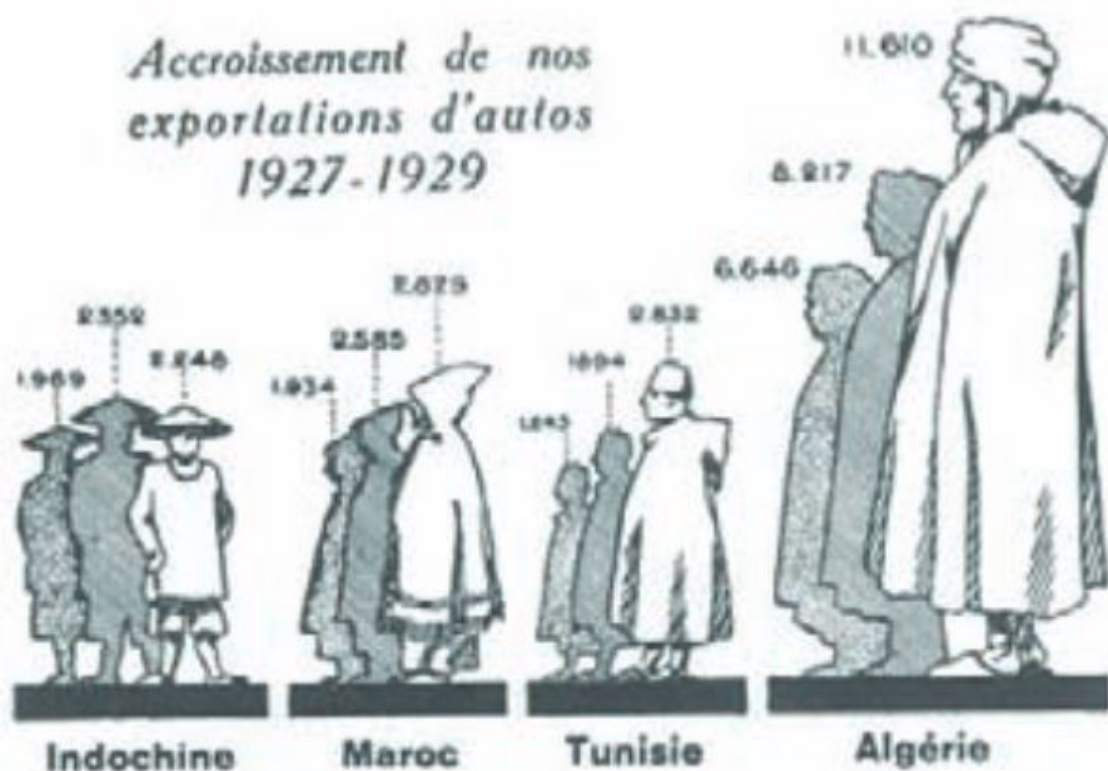


Figura 13.1: Fuente: Satet, R. (1932), *Les Graphiques*, Paris. Incluido en [Tufte 2001](#)

Al observarla Figura 13.1, nuestra mente compara el tamaño relativo de cada objeto. Los valores relativos están distorsionados en dos sentidos:

1. La escala no es exacta. La altura a la que aparecen las cifras de Argelia debe ser mucho mayor. Un fenómeno cuatro veces mayor, debe ser mostrado cuatro veces más grande o cuatro veces más alto.



2. Al superponer la figura referente al año más reciente sobre las de los años anteriores, la diferencia de valor entre los años es más difícil de calcular, pues las figuras en primer plano parecen mucho más grandes que las del fondo. Sólo la altura relativa nos indica el cambio ocurrido entre los años.

### **La percepción también se basa en la experiencia**

La experiencia también juega un papel en cómo se perciben los gráficos. Los estadísticos utilizan elementos visuales para explicar los resultados, debido a que un público más amplio necesita una mayor ayuda para entender la información estadística. Una presentación visual de los datos debe hacer las conclusiones principales fáciles de apreciar y entender.

### **No hacer un uso indebido de las herramientas**

Hoy en día se dispone de numerosas herramientas tecnológicas enfocadas a presentar de manera visual información numérica. Hay que asegurarse de que la atención se centra en el contenido y el mensaje de la presentación gráfica, más que en la metodología, el diseño o la tecnología utilizada.

### **Elementos a tener en cuenta para una correcta representación visual de los datos**

Cuando se realizan representaciones visuales, se debe tener en cuenta los siguientes factores:

- El grupo-objetivo: diferentes formas de presentación pueden ser adecuadas para diferentes públicos (por ejemplo, de negocios o del mundo académico, especialistas o población general).
- El papel de la gráfica en la presentación: analizar el cuadro completo o centrar la atención en puntos clave que pueden requerir diferentes tipos de representaciones visuales.
- Cómo y dónde será presentado el mensaje: un análisis largo y detallado o una presentación rápida.
- Aspectos contextuales que pueden distorsionar la comprensión: usuarios de datos expertos o novatos.
- Si es mejor solución el análisis textual o una tabla de datos.
- Consideraciones sobre accesibilidad: \* Ofrecer alternativas de texto para los elementos no textuales como gráficos e imágenes; \* No depender únicamente del color. Si prescindimos del color, ¿la presentación sigue siendo comprensible? ¿Tienen las combinaciones de colores suficiente contraste? ¿Son válidos esos colores para daltónicos (rojo / verde)?; \* Asegurar que el contenido de reproducción puede ser controlado por el usuario (por ejemplo, pausando gráficos animados).
- Coherencia entre las representaciones visuales de datos: asegurar que los elementos de las representaciones visuales estén diseñados de forma coherente y

que usa las convenciones habituales cuando sea posible (por ejemplo, azul para representar el agua en un mapa).

- Tamaño, duración y complejidad: ¿Es la presentación fácil de entender? ¿Es demasiado pesada/larga/compleja para el público en una sesión dada?
- Posibilidad de una mala interpretación: probar la presentación con los compañeros, amigos o alguna persona del grupo objetivo, para ver si se capta el mensaje pretendido.

## 13.4 Tablas

### Por qué las tablas son importantes

Unas tablas adecuadas son parte esencial del conjunto, ya se trate de una nota de prensa, un artículo de análisis o de un trabajo de investigación. Utilizar tablas de forma efectiva ayuda a minimizar el número de datos en el texto.

### Dos tipos de tablas

Existen dos tipos de tablas. Las primeras son las tablas más pequeñas, llamadas *tablas de presentación* (o de demostración). Se usan, principalmente, para resaltar los elementos principales en una nota de prensa, página web o publicación de análisis.

El segundo tipo consiste en tablas más grandes, llamadas *de referencia*. Éstas están siendo reemplazadas progresivamente por bases de datos interactivas, que permiten a los usuarios generar en línea sus propias tablas.

Cada tabla debe contener suficientes metadatos, como un título descriptivo y una indicación de la fuente de procedencia, para que pueda ser copiado y pegado en otro documento sin perder su sentido.

### Elementos a tener en cuenta para el buen diseño de una tabla

Veamos cinco elementos de apoyo que son necesarios para la descripción de los datos mostrados en una tabla:

- El título de la tabla debe hacer una descripción clara y precisa de los datos. Debe responder las tres preguntas 'qué', 'dónde' y 'cuándo'. Se debe ser breve y conciso, y evitar el uso de verbos.
- Los encabezados de las columnas, expuestos en la parte superior de la tabla, deben indicar qué datos hay presentes en cada columna de la tabla y proporcionar los metadatos necesarios (por ejemplo, unidad de medida, período de tiempo o área geográfica).
- Los encabezados de las filas, en la primera columna de la tabla, deben identificar los datos presentes en cada fila de la tabla.

- Las notas a pie, en la parte inferior de la tabla, pueden proporcionar cualquier información adicional necesaria para comprender y utilizar correctamente los datos (definiciones, por ejemplo).
- La línea que contiene la fuente, en la parte inferior de la tabla, debe indicar la fuente de los datos, es decir, la organización que elaboró los datos y el método de recogida de datos (por ejemplo, censo de población o encuesta de población activa).

La Figura 13.2 indica cómo mostrar de manera adecuada estos componentes en una tabla.

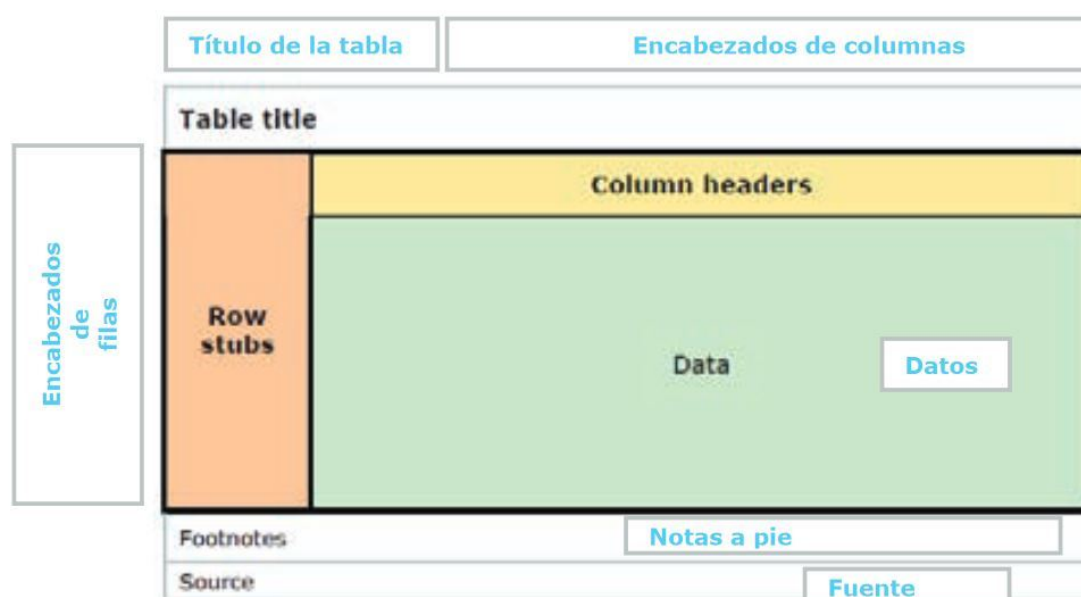


Figura 13.2: Componentes de una tabla.

Para asegurar que las tablas son fáciles de entender, se deben considerar las siguientes cuestiones:

- Evitar texto innecesario.
- Presentar los datos por orden cronológico en el caso de series temporales o usando una clasificación estándar. Para series temporales más largas, puede ser más apropiado utilizar el orden cronológico inverso (es decir, comenzando con el período más reciente y yendo hacia atrás), como cuando se considera el desempleo mensual.
- Usar un número reducido de decimales.
- Utilizar separadores de miles. El uso de un espacio en lugar de un símbolo puede evitar el problema de tener que traducir entre diferentes lenguas.
- Alinear los números hacia la coma decimal (o a la derecha en la ausencia de cifras decimales) hace claramente apreciable su valor relativo. No centrar los números en una columna, a menos que todos ellos tengan la misma extensión.

- No dejar ninguna celda de datos vacía. Los valores que faltan deben ser identificados como 'Dato no disponible por secreto estadístico' o 'Dato no aplicable'.

### El uso de redondeo y decimales

El redondeo se puede utilizar cuando los datos no requieren de un alto grado de precisión. Si se necesita mostrar los valores con un número de decimales variable, se deberían alinear en el punto decimal, no a la derecha. Los valores numéricos deben justificarse a la derecha.

Buen ejemplo	Mal ejemplo
93.2	93.2
1045.0	1045
385.6	385.62

## 13.5 Gráficos

### ¿Por qué usar gráficos?

Las estadísticas a menudo pueden entenderse mejor si se presentan en un gráfico en lugar de en una tabla. Un gráfico es una representación visual de los datos estadísticos, en el que los datos están representados por símbolos como barras o líneas. Es una herramienta visual muy eficaz, ya que muestra datos de manera rápida y sencilla, facilita la comparación, y puede revelar las tendencias y las relaciones entre los datos.

En general, un gráfico adopta la forma de una figura en una o dos dimensiones, como un gráfico de barras o un gráfico de líneas. También se dispone de gráficos en tres dimensiones aunque, por lo general, se consideran demasiado complejos de entender.

Los gráficos se pueden utilizar para ilustrar patrones de grandes cantidades de datos o para comunicar un hallazgo clave o un mensaje.

### Lista de elementos para diseñar un buen gráfico

Un buen gráfico capta la atención del lector; presenta la información de forma sencilla, clara y precisa; no induce a error; resume la información de los datos; facilita la comparación de datos y destaca las tendencias y las diferencias; ilustra el mensaje, tema o trama del texto a que acompaña.

### Cuándo puede no ser apropiado utilizar gráficos

Un gráfico no es siempre la herramienta más apropiada para presentar la información estadística. A veces un texto y/o tabla de datos puede proporcionar una mejor explicación al público permitiendo ahorrar mucho tiempo y esfuerzo. Se debe reconsiderar el uso de gráficos cuando los datos: son muy dispersos; tienen muy pocos valores; tienen

demasiados valores; guardan poca o ninguna variación.

### Seleccionar el tipo de gráfico apropiado

Saber qué tipo de gráfico usar para cada tipo de información es crucial. Algunos gráficos son más apropiados que otros, dependiendo de la naturaleza de los datos.

- **Gráfico de barras**

El gráfico de barras mostrado en la Figura 13.3 es el tipo de gráfico más simple de dibujar y leer. Se utiliza para comparar las frecuencias o los valores de las distintas categorías o grupos. Las barras pueden ser tanto verticales como horizontales. En la orientación horizontal, el texto es más fácil de leer. También es algo más fácil comparar los diferentes valores cuando las barras están ordenadas del tamaño más pequeño al más grande, en lugar de estar dispuestas arbitrariamente.

Un gráfico de barras apiladas se puede utilizar para mostrar y comparar segmentos dentro de unos totales. Se debe tener precaución con el uso de este tipo de gráficos, ya que pueden ser difíciles de analizar y comparar si hay demasiados elementos en cada pila o si muchas barras son de un tamaño similar.

Una pirámide de población es una combinación de dos gráficos de barras horizontales, representando la estructura por edades de la población femenina y masculina de un país o región. Convencionalmente los hombres se muestran a la izquierda y las mujeres a la derecha. Cuando se quieren comparar diferentes pirámides de población, por lo general es mejor representar el porcentaje de hombres y mujeres sobre el total de la población, en lugar de su número.

- **Gráficos de líneas**

El gráfico de líneas en la Figura 13.4 es una herramienta eficaz para la visualización de tendencias en los datos a lo largo del tiempo y, por lo tanto, el tipo de gráfico más adecuado para series temporales. Los parámetros del gráfico se pueden ajustar para comunicar mejor el mensaje, pero se debe tener cuidado de no distorsionar los datos.

- **Gráficos circulares**

Un gráfico circular de la Figura 13.5 se puede utilizar para mostrar la distribución porcentual de una variable, pero sólo se puede mostrar un pequeño número de categorías (segmentos), por lo general no superior a seis. Muchos estadísticos desaconsejan el uso de este tipo de gráficos, ya que puede ser difícil comparar los diferentes segmentos del círculo y, aún más, comparar datos entre diferentes gráficos circulares. Para evitar este problema, los segmentos pueden ser etiquetados con sus valores reales. En algunos casos, los nombres de las categorías también se

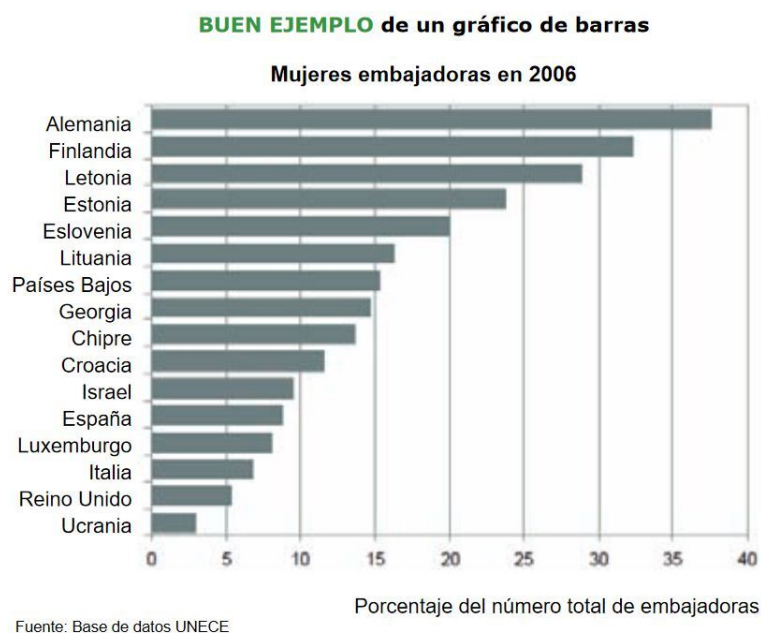


Figura 13.3: Gráfico de barras.

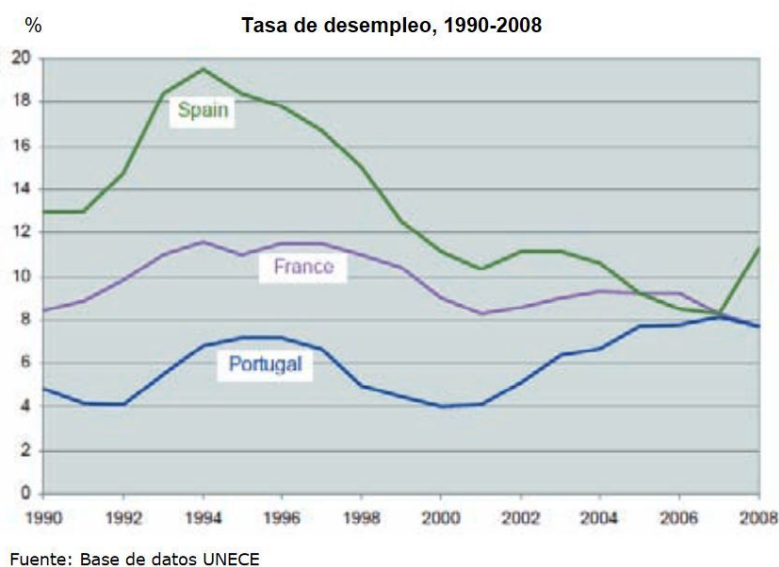


Figura 13.4: Gráfico de líneas.

pueden escribir como etiquetas en el gráfico, de modo que no resulte necesaria añadir una leyenda. Los segmentos, en general, como están mejor presentados es en orden de menor a mayor, en lugar de intercalar segmentos pequeños y grandes.

- **Diagrama de dispersión**

El gráfico de dispersión se utiliza para mostrar la relación entre dos variables. Es la manera más exacta de comprobar si existen correlaciones.

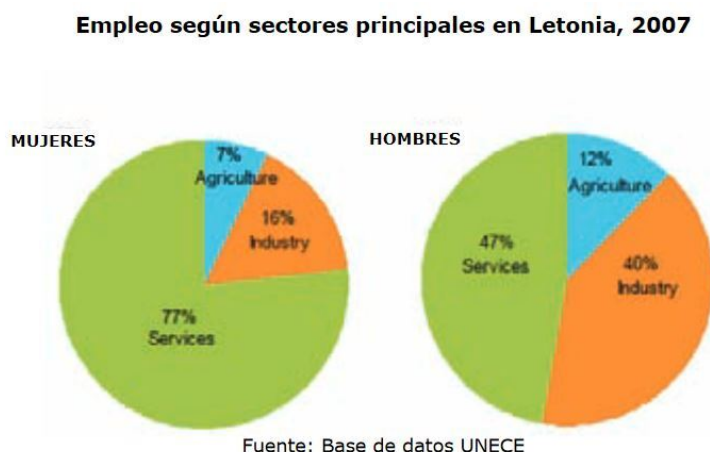


Figura 13.5: Gráfico circular.

### ¿Qué hace a un gráfico efectivo?

Entre los componentes de un gráfico están:

1. Componentes de representación de los datos: barras, líneas, áreas o puntos.
2. Componentes de apoyo que ayudan a la comprensión de los datos. Estos son:
  - El título del gráfico, que debe dar una idea clara acerca de lo que trata el gráfico. Tiene que ser corto y conciso. Se pueden utilizar dos tipos de títulos: - Un título informativo proporciona toda la información necesaria para entender los datos; - Un título descriptivo pone de relieve las principales pautas o tendencias representadas en el gráfico y plantea en pocas palabras aquello que ilustra el gráfico.
  - Las etiquetas de los ejes, que deben identificar los valores mostrados en el gráfico. Los niveles de las categorías se muestran horizontalmente en ambos ejes.
  - Los títulos de los ejes, que deben identificar la unidad de medida de los datos (por ejemplo, 'en miles', '%', 'edad (en años)' o '\$'). No es necesario que incluya un título de eje cuando la unidad de medida es evidente (por ejemplo, 'años' para series temporales).
  - Las líneas de división, que se pueden añadir en los gráficos de barras y de líneas para ayudar a los usuarios a leer y comparar los valores de los datos.
  - Las leyendas y etiquetas de datos, que deben identificar los símbolos, patrones o colores utilizados para representar los datos en el gráfico. Cuando en el gráfico se representa únicamente una serie de valores, no se debe mostrar la leyenda. Es más, siempre que sea posible se deben utilizar etiquetas de datos en lugar de leyendas. Las etiquetas de datos se muestran en (o junto a) los componentes de representación para facilitar su identificación y

comprensión.

- Las notas a pie pueden ser utilizadas para proporcionar definiciones o información metodológica.
- La fuente de los datos debe estar indicada en la parte inferior de la tabla.

### Ajustar los parámetros del gráfico

Al diseñar un gráfico, se puede ajustar su escala para una mejor transmisión del mensaje. Una buena práctica es utilizar algún símbolo para indicar cuándo una escala de valores no comienza en cero o cuando en algún momento se deja de mantener la escala. La mejor opción es empezar de cero y poner ya sea una línea en zigzag o un hueco indicando que se ha dejado de mantener la escala. Si nos fijamos en la Figura 13.6, vemos que la longitud de la barra correspondiente al año 2020 no mantiene la proporción con respecto al resto. Para alertar de esto, se recurre a la representación usando la línea gruesa en zigzag. Este recurso permite que en un mismo gráfico se puedan ver las barras correspondientes a años con valores más pequeños que, en el caso de mantener rígidamente las proporciones, no se verían con claridad.

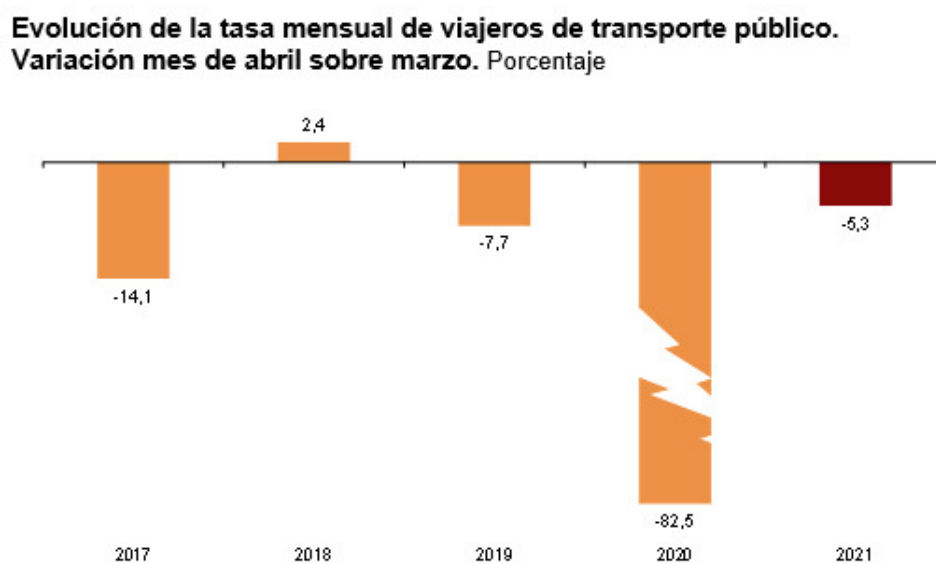


Figura 13.6: Ajuste de parámetros.

### El control de la carga cognitiva de los gráficos

Carga cognitiva significa básicamente cuánto tiene que esforzarse el lector para entender lo que se está tratando de comunicar. Un gráfico con una elevada carga cognitiva será difícil de entender y recordar y su mensaje será difícil de comunicar. Un gráfico con una reducida carga cognitiva se entiende con un vistazo y su mensaje será obvio. Cuando se diseñe un gráfico, controlar la carga cognitiva. Se puede reducir y enviar un mensaje claro, usando las convenciones y formatos adecuados.



### Sugerencias para mejorar los gráficos

- Ser preciso: Los elementos gráficos deben ser de un tamaño tal que represente los índices con precisión, véase Figura 13.7.

#### MAL EJEMPLO de tamaño relativo entre objetos gráficos

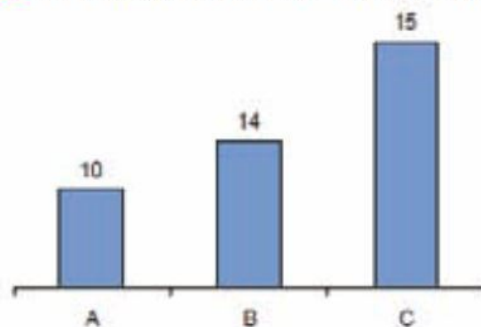


Figura 13.7: Importancia de la precisión.

- Ordenar los datos: Cuando se usan gráficos de barras o de pastel, se deben ordenar los datos de menor a mayor valor, para que resulten más fáciles de comparar. Ver Figura 13.8.

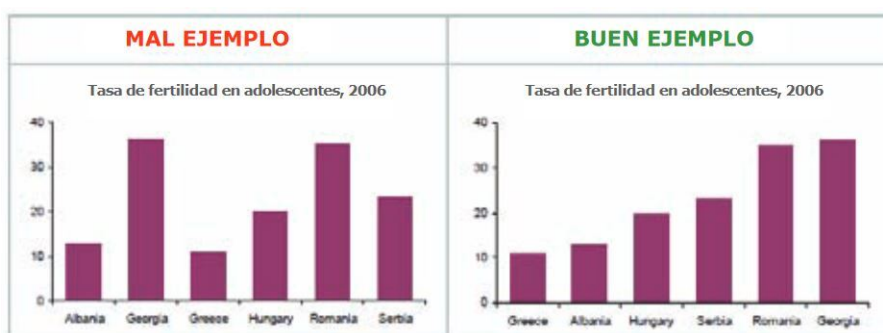


Figura 13.8: Datos ordenados.

- Evitar correlaciones erróneas: Trazar variables con diferentes escalas en el mismo gráfico puede conducir a conclusiones erróneas. El hecho de que dos curvas se muevan juntas no es suficiente para establecer una correlación.
- Utilizar con precaución un doble eje Y: Un doble eje Y puede generar confusión. Este gráfico es útil cuando se tienen dos variables diferentes. Pero se debe tener mucho cuidado con las etiquetas y con mostrar cada línea de datos en el mismo color que el eje que el usuario necesita consultar.
- Evitar elementos innecesarios en el gráfico: Los gráficos en tres dimensiones rara vez añaden valor y a menudo confunden a los lectores. En general, se debería evitar añadir distintos colores en los gráficos de barras como se ilustra en la Figura 13.9.

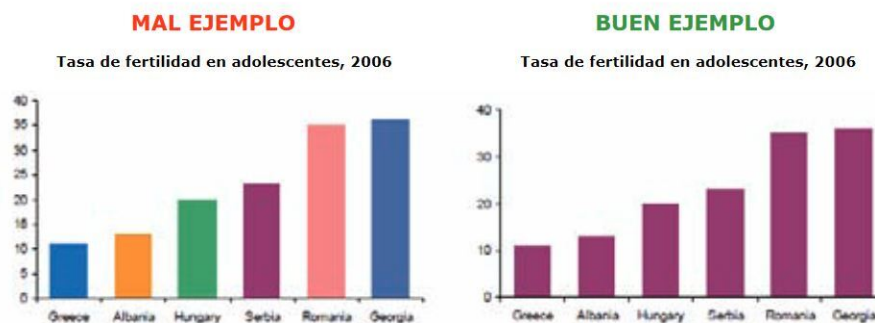


Figura 13.9: Gráficos de barras de distintos colores.

## 13.6 Mapas

### Por qué un mapa es mejor que miles de números

La información geográfica es una parte esencial de todo dato estadístico. Las áreas geográficas tienen límites, nombres y otra información que permite su localización sobre el terreno y relacionar con ello la información estadística. Los mapas temáticos (estadísticos) se utilizan para mostrar la distribución espacial de uno o más atributos estadísticos. Los mapas son las herramientas más eficientes para visualizar los patrones espaciales.

Un mapa está bien diseñado si es fácil de entender, si puede resumir voluminosas tablas de datos o textos largos y complicados.

Los mapas pueden ser muy útiles tanto en la preparación de los censos y encuestas, como en el análisis y presentación de informes de resultados. Se debe considerar el uso de mapas si desea:

- mostrar la localización geográfica y la distribución espacial de los datos;
- comparar las diferentes áreas;
- resumir un gran volumen de datos y reducir su complejidad;
- comunicar un mensaje claro;
- confirmar resultados obtenidos;
- atraer la atención de la gente;
- acumular información espacial en un sistema de información geográfica.

### Elementos a tener en cuenta para el buen diseño de un mapa

Un buen mapa:

- es simple y fácil de entender;

- tiene un mensaje claro y objetivo;
- proporciona una representación exacta de los datos y no induce a error;
- atrae la atención del lector hacia la información más importante;
- está bien presentado y resulta atractivo;
- hace concordar el formato de presentación y el público;
- se puede sostener por sí mismo sin requerir más explicaciones;
- es accesible para las personas daltónicas.

### **Cuándo carece de sentido hacer un mapa**

No tiene sentido realizar un mapa si:

- los datos no tienen desglose geográfico;
- no hay variaciones significativas entre los datos;
- el público objetivo puede tener dificultades para entender el mapa;
- no hay suficiente espacio disponible para presentar el mapa de forma que pueda ser leído y entendido correctamente.

### **Mapas estáticos o interactivos**

Los mapas pueden ser estáticos o interactivos. Los mapas estáticos no pueden ser editados por el usuario. Los mapas interactivos ofrecen flexibilidad y dan al usuario la capacidad de modificar el diseño, seleccionar y recuperar datos, dotar de animación el mapa, y cambiar los temas o centrar la atención sobre aspectos considerados de interés fundamental.

### **Seleccionando el tipo apropiado de mapa**

El mismo consejo para los gráficos se aplica a los mapas: es fundamental saber qué tipo de mapa se va a producir y qué tipo de información. La selección de la técnica apropiada de trazado de mapas, depende de la naturaleza de los datos.

- **Mapas de coropletas**

El tipo más común de mapa es el mapa de coropletas, en el que las áreas se colorean en función del valor de la variable que se muestra, como en la Figura 13.10. Este tipo de mapa proporciona una manera fácil de visualizar los patrones a través del espacio. Los mapas de coropletas deben ser utilizados preferentemente para mostrar los fenómenos que se distribuyen uniformemente dentro de cada unidad espacial.

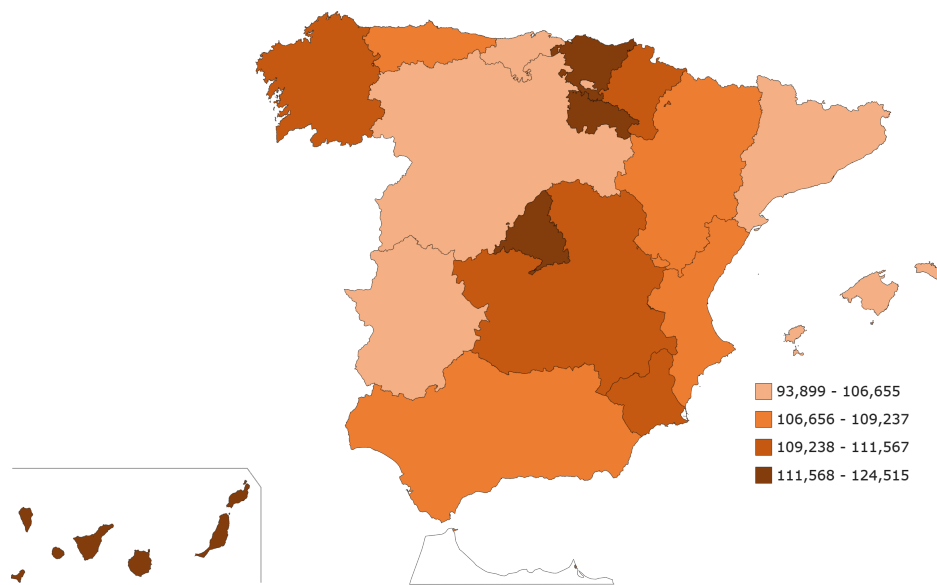


Figura 13.10: Índices de Comercio al por Menor, Comunidades y Ciudades Autónomas, Índice General, 2019M01. Fuente: INE

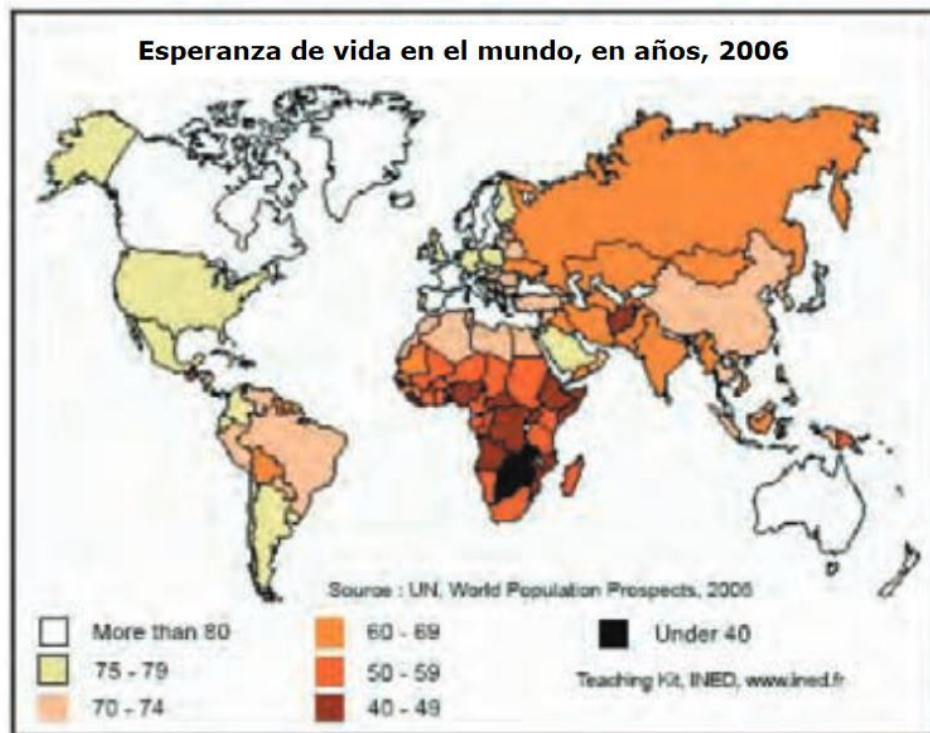
- **Mapas de puntos** Un mapa de puntos muestra la ubicación y la densidad de una población o fenómeno usando símbolos. Permite a los usuarios captar rápidamente la magnitud general de los datos, así como su concentración o dispersión. Cada punto representa un valor discreto, por lo general un gran número de entidades.
- **Mapas de símbolos proporcionales** Un mapa de símbolos proporcionales (o graduados) se utiliza para mostrar valores absolutos. El tamaño del símbolo es proporcional al tamaño de la población o fenómeno que se representa. Cada símbolo está unido a un punto específico dentro de la unidad espacial, por lo general el centro o la capital.

### Consejos de diseño

Como con cualquier método de visualización de datos, la recomendación más importante para asegurar que el mensaje se haga entender, es hacer el mapa simple.

Los mapas se deben diseñar para que sean independientes del texto relatado o de las tablas de datos. Al igual que un gráfico, un mapa debe entenderse por sí solo, sin necesidad de texto o notas a su alrededor de manera que, una vez publicado, pueda ser escaneado o descargado y utilizado en otro contexto. Ver Figura 13.11.

Los componentes de un mapa compiten entre sí por la atención del lector. Para maximizar la eficacia del mensaje, debemos asegurarnos de que los datos son el centro del diseño. Los siguientes componentes son necesarios para ayudar al usuario a entender el mapa:



Fuente: Institut National d'Etudes Démographiques (INED)

Figura 13.11: Mapa que incluye toda la información.

- El título del mapa, que debe dar una idea clara acerca de qué es lo que trata el mapa. Tiene que ser corto y conciso. Los subtítulos pueden añadirse para proporcionar información más detallada (por ejemplo, unidad de medida).
- La leyenda, que debe identificar todos los símbolos, patrones y colores utilizados para representar los datos en el mapa.
- Las unidades geográficas en las que los datos están representados en el mapa, que deben estar identificadas, ya sea en el título (o subtítulo) o en una leyenda.
- Las etiquetas de texto pueden ser añadidas en el mapa para identificar los lugares importantes o relevantes u otra información.
- La escala del mapa, que puede ser proporcionada para ayudar al usuario a medir distancias y comparar diferentes mapas.
- Una nota a pie, que puede ser utilizada para proporcionar definiciones u otra información metodológica.
- La fuente de los datos, que debe ser identificada en la parte inferior del mapa.
- En la información de copyright se debe identificar al autor responsable del contenido, exponiéndolo en la parte inferior del mapa.

Otros componentes, que no son esenciales pero que pueden ser incluidos en algunos mapas son:

- Una flecha indicando el norte, que solo es necesaria cuando el mapa no está orientado hacia el norte.
- Latitudes y longitudes, que sólo se necesitan en mapamundis o mapas continentales.
- Un mapa de ubicación, que es una pequeña réplica del mapa base, que sitúa el área asignada en su contexto más amplio. Puede ser útil si su público no está familiarizado con la geografía de la zona.
- Otros gráficos, que se pueden añadir si mejoran la comprensión del mapa.

La redacción del título y la leyenda deben ser concisos, usar solamente palabras clave y evitar la repetición de las mismas palabras en el título, leyendas o pies de página, evitar abreviaturas y acrónimos, utilizar un tamaño de fuente más pequeño en las leyendas que en el título y una fuente aún más pequeña para el pie de página.

El color es una de las características gráficas más potentes. En primer lugar, debes estar al tanto de cualquier convencionalismo existente asociado a cualquier color elegido, así como las posibles connotaciones positivas o negativas.

Después, debes asegurarte de que cualquiera puede comprender el mensaje con los colores utilizados en el mapa. Por ejemplo, los daltónicos tienen dificultades para distinguir entre algunos colores. El caso más común es el rojo-verde.

Por último, cuando hay relativamente pocas clases de datos para los valores en una escala continua (por ejemplo, la densidad de población), debes considerar el uso de diferentes tonalidades del mismo color en vez de colores diferentes.

## 13.7 Técnicas de visualización emergentes

### ¿Por qué una representación es más que una imagen?

Las herramientas y técnicas emergentes están proporcionando nuevas oportunidades para la visualización de datos y para hacerlos más interesantes a los usuarios. Los generadores de tablas dinámicas, gráficos y mapas, permiten a los usuarios manipular los datos y crear sus propias representaciones.

#### Visualizaciones dinámicas

Es muy común disponer de un conjunto de herramientas de visualización que permiten a los usuarios crear tablas, gráficos o mapas en línea, sin tener que descargar los datos y trabajar con otra aplicación.

## **Animación y vídeo**

La animación y el vídeo son dos importantes técnicas emergentes de visualización de datos. Este formato hace más fácil contar la historia, mediante la combinación de descripciones en audio o texto con ilustraciones gráficas que explican el significado que se esconde tras los números.

## **13.8 Cuestiones de accesibilidad**

Para que la información estadística básica sea utilizada de forma amplia y sencilla, debe ser universalmente accesible. Esto significa que cualquiera debe ser capaz de consultar y entender la información, independientemente de la tecnología que utilice o de cualquier discapacidad que pueda tener.

El principio de igualdad de acceso a la información estadística básica es muy importante. Una estrategia de accesibilidad bien pensada beneficiará a todos.

### **Texto**

El texto debe ser conciso, coherente y estar bien estructurado para que los usuarios puedan encontrar fácilmente la información que buscan. Presentando el texto en secciones lógicas y bien diferenciadas con títulos y subtítulos se hará más sencillo realizar búsquedas en él y la conversión a otros formatos.

Para servir con eficacia a todos los grupos objetivo, el texto debe estar disponible en múltiples formatos, por ejemplo en Braille, audio o letras de gran tamaño.

### **Tablas**

Cuando se utilicen tablas para presentar los datos, también se debe prestar atención a la accesibilidad. Si los datos se exponen sin toda la información necesaria para comprenderlos e interpretarlos, estos resultarán inútiles o engañosos.

Si una tabla se extiende a lo largo de varias páginas es importante repetir los encabezados de columna y de fila en cada nueva página.

En formatos electrónicos, tales como HTML, se puede incluir las etiquetas para los encabezados de columna y fila, junto con una breve descripción de su contenido cuando sea necesario. El uso de etiquetas es bueno para todos.

### **Gráficos**

Cuando se generan gráficos, merece la pena tener en cuenta que no todos los usuarios tienen acceso a imágenes. Se puede realizar una descripción de texto para proporcionar la misma información que se muestra en el gráfico. Estas descripciones también

pueden ser utilizadas en la producción de documentos en audio, Braille y otros formatos.

## Mapas

Los mapas constituyen un desafío técnico muy importante en términos de accesibilidad para las personas con alguna discapacidad. La información se transmite mediante una combinación de imágenes y colores, dos métodos prácticamente incompatibles con los estándares de accesibilidad. Por tanto, debemos pensar en ofrecer un texto alternativo para proporcionar la misma información que se representa en el mapa y/o dar acceso a las tablas de datos.

## Metadatos

Las organizaciones de Estadística deberían garantizar que los usuarios dispongan de los metadatos que necesiten para comprender los datos, incluyendo sus puntos fuertes y sus limitaciones. Estos metadatos, deben mantenerse al día mediante la incorporación de los últimos cambios en las definiciones, clasificaciones y metodología. También deben informar sobre el estado de los datos (si son provisionales o definitivos).

Por último, se debe tener en cuenta las siguientes recomendaciones:

- Todos los metadatos deberían estar disponibles en Internet de forma gratuita.
- Presentar los metadatos de tal manera que respondan a las necesidades de una amplia gama de usuarios con diferentes necesidades y/o conocimientos estadísticos.
- Asegurar que los vínculos a los metadatos de los cuadros y gráficos que están describiendo están activos y viceversa.
- Disponer de los metadatos no sólo en la lengua nacional sino también, si es posible, en una lengua de uso común como el inglés.
- Proporcionar un motor de búsqueda local, basado en la búsqueda de texto libre.

## Bibliografía

Tufte, E.R. (2001). *The Visual Display of Quantitative Information*. 2nd. Cheshire CT, Graphics Press.

UNECE (2009). *Cómo hacer comprensibles los datos. Parte 2: Una guía para presentar estadísticas*. URL: [https://unece.org/DAM/stats/documents/writing/MDM\\_Part2\\_Spanish.pdf](https://unece.org/DAM/stats/documents/writing/MDM_Part2_Spanish.pdf).



## Tema 14

### **Record linkage. Introducción. Visión de conjunto de de los métodos. Preparación de los datos**

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

W.E. Winkler (2009). *Record linkage*, en D. Pfeferman and C.R. Rao (eds.), *Handbook of Statistics 29A*, cap. 14, pp. 351-370. Amsterdam: North-Holland

M. Pérez Julián A. Fresneda Pacheco (2014). *Use of administrative records in official statistics: a practical view*. URL: <http://www.seio.es/BEIO/Use-of-administrative-records-in-official-statistics-a-practical-view-2.html>

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.

### **14.1 Introducción**

El *record linkage* consiste en un conjunto de métodos para emparejar (*matching*) duplicados dentro de o entre ficheros usando identificadores no exclusivos como el nombre, el apellido, la fecha de nacimiento, el D.N.I., la razón social, el N.I.F., la dirección, municipio/ provincia/Comunidad Autónoma de residencia de la persona o de ubicación de la empresa/establecimiento y otras características. Nos referiremos a los campos como el nombre, el apellido, la fecha de nacimiento, el D.N.I., el N.I.F. o la dirección como *cuasi-identificadores*. Combinándolos, los cuasi-identificadores pueden identificar a una unidad física o jurídica de forma única.

Los métodos modernos de *record linkage* se empezaron a usar en el campo de la genética a principios de los 60 por Howard Newcombe, usando razones de probabilidad y probabilidades de valores específicos basados en la frecuencia. Fellegi y Sunter 1969, a finales de los 60, le dieron una formalización matemática a estas ideas. Probaron la

optimalidad de la regla de decisión/clasificación de Newcombe e introdujeron muchas ideas sobre los parámetros óptimos de estimación sin un conjunto de datos de entrenamiento ([Winkler 2006](#)).

Aunque los métodos se basan en modelos estadísticos, la mayoría de los desarrollos han sido hechos por informáticos, usando principalmente machine learning sobre bases de datos.

Las aplicaciones del *record linkage* son numerosas. Una primera aplicación es la creación de una lista o de un marco muestral a partir de una colección de listas más pequeñas en el marco de una encuesta o actualizar una lista ya existente. La actualización y el mantenimiento de listas nos asegurará que tenemos una buena cobertura de una población y evitará problemas de cobertura de marcos. Los errores tipográficos en campos como el nombre, los apellidos, la razón social, la dirección o la fecha de nacimiento en un conjunto de registros puede hacer que la actualización se complique. Históricamente estos trabajos de depuración de listados se llevaban a cabo de forma manual. Los métodos informáticos de *record linkage* pueden reducir significativamente la necesidad del trabajo manual, con el consiguiente ahorro de tiempo y dinero.

Otra aplicación del *record linkage* es el cruce de una lista con otra para estimar la subcobertura/ sobrecobertura de una de las listas, que se considera bastante completa. Por ejemplo, en el caso de obtención de hogares a partir de individuos.

Como vemos, el *record linkage* se puede usar tanto para mejorar la cobertura como para reducir los duplicados en el marco de una encuesta. Esto es muy importante ya que los errores de marco pueden afectar al sesgo y por tanto a las estimaciones obtenidas a partir de ella. Es casi imposible corregir el error de estimaciones que se basan en un muestreo con un marco que contenga errores.

En este tema veremos el origen del modelo de Fellegi y Sunter de *record linkage*, los métodos de estimación de parámetros sin un conjunto de datos de entrenamiento y comparadores de cadenas para tratar los errores tipográficos en la Sección [14.3](#). La Sección [14.4](#) incluye información sobre las dificultades al trabajar con datos desordenados, la preparación de los ficheros de datos puede ayudar a mejorar la eficacia de los métodos. Como el *record linkage* se usa mucho con datos administrativos, veremos qué son y cómo se usan en la estadística oficial los registros administrativos (Sección [14.2](#)) y por último cómo usar el *record linkage* con datos administrativos (Sección [14.5](#)).

## 14.2 Los datos administrativos en la estadística oficial

En la estadística oficial existen distintas fuentes de datos, las principales son las encuestas, los censos y los datos administrativos. Los datos administrativos<sup>1</sup> son el conjunto de unidades y datos obtenidos de una fuente administrativa<sup>2</sup>. Este tipo de datos se recogen con el fin de llevar a cabo actividades no estadísticas y del mantenimiento de estos registros se encargan organismos que dependen del gobierno u organizaciones privadas. Por ejemplo, los registros administrativos se utilizan para regular el flujo de bienes y de personas en las fronteras, para responder a necesidades legales de registrar eventos particulares como los nacimientos y las muertes, y para administrar prestaciones como pensiones u obligaciones como los impuestos de personas y empresas.

Los datos administrativos son recogidos por unidades administrativas e informan sobre variables administrativas. Una de las principales ventajas es su bajo coste, además de la reducción de la carga al informante. Pero su uso para fines estadísticos depende esencialmente de los metadatos, que proporcionan información detallada sobre las definiciones usadas, las características de los datos, el contexto institucional, o los procedimientos que se usan para fines administrativos.

Los registros, que hace años se guardaban en papel, en la administración pública han sido reemplazados por ficheros de datos digitalizados. Esta digitalización de los registros, además del desarrollo del *record linkage*, han propiciado el uso cada vez más extendido de los datos administrativos con fines estadísticos.

Actualmente, el uso de datos administrativos con fines estadísticos es una práctica común. Ésta no debe entenderse exclusivamente desde la perspectiva de la producción estadística de operaciones estadísticas usando registros en lugar de encuestas, sino como una herramienta adicional que complete la producción estadística y permita llevar a cabo procesos mixtos en los cuales tanto las encuestas como los registros administrativos puedan ser usados. La combinación de ambos en el proceso estadístico permite la reducción del presupuesto y de la carga al informante, disponer de información auxiliar para el uso de estimadores de razón o para la imputación de la falta de respuesta.

Tal y como se indica en [A. Wallgren y B. Wallgren 2007](#), algunos de los usos de los datos administrativos son:

- en la producción estadística sustituyen los datos de una encuesta;
- como input para los registros estadísticos que serán usados como marco muestral y como fuente de información auxiliar en el diseño muestral;
- como fuente de variables adicionales en el uso de estimadores, por ejemplo, para el estimador de razón;

---

<sup>1</sup>[https://ec.europa.eu/eurostat/cros/content/administrative-data-0\\_en](https://ec.europa.eu/eurostat/cros/content/administrative-data-0_en)

<sup>2</sup>Una fuente Administrativa es un repositorio de datos que contiene información recogida y mantenida con el fin de implementar una o más normas administrativas.

- como información auxiliar en algunas fases del proceso estadísticos (por ejemplo, en la depuración de datos, en la imputación, en la calibración de estimaciones).

Para que el uso de *record linkage* en la estadística oficial aplicado a los datos administrativos sea lo más satisfactorio posible es necesario que los registros administrativos verifiquen las siguientes características:

- a. Debe incluir información sobre los fenómenos estudiados referidos al total de la población que queremos analizar.
- b. Esta información debe estar actualizada a la fecha de referencia del estudio.
- c. El registro debe contener algún identificador que permita la identificación de cada ítem (persona, vivienda, empresa, etc.) del acto administrativo que refleja, permitiendo el cruce con otras fuentes de forma que se pueda obtener toda la información necesaria para el proceso estadístico.
- d. Es obligatoria la ausencia total de duplicados y registros inválidos (por ejemplo, unidades vacías, o que aparezcan en el listado unidades que no debieran estar) ya que distorsionan el marco de la población de referencia y puede producir sesgos en las estimaciones.

Lamentablemente, en términos de conceptos, las definiciones administrativas no siempre coinciden con las estadísticas. En muchas operaciones estadísticas las definiciones vienen fijadas por ley y se establecen en organismos internacionales con el fin de que los datos sean comparables a nivel europeo y mundial, mientras que los registros administrativos tienen que satisfacer necesidades de los gobiernos de cada país. Además, cambios en las necesidades del gobierno puede provocar cambios en los conceptos administrativos que pueden hacer que las series de datos temporales tengan cortes o que haya que hacer modificaciones para poder utilizar los datos con fines estadísticos, por ejemplo, si los datos administrativos incluyen el IVA y es necesario su eliminación para su uso estadístico, o viceversa.

El uso de registros administrativos junto con otras fuentes de datos, como pueden ser las encuestas, conlleva a que los errores que se pueden cometer en el proceso estadístico aumenten. [Statistics New Zealand 2016](#) ha creado un marco de error para evaluar la calidad del uso de los registros administrativos en las oficinas estadísticas. El modelo del 'Total Survey Error' desarrollado por Groves et al., ver Figura 14.1, examina todas las posibles fuentes de error en las encuestas. En 2012, Li-Chun Zhang desarrolló un nuevo marco de error que descompone los pasos entre los conceptos ideales y la población que nos gustaría reflejar en nuestros ficheros de datos y en los datos finales que obtenemos en la práctica. Mediante el uso del marco de Li-Chun Zhang podemos compilar una lista exhaustiva de las fuentes de datos para un conjunto de datos dados. Este marco de error separa el 'ciclo de vida' de los datos estadísticos en dos fases lo que hace más fácil categorizar las fuentes de error y entender sus causas. La idea es evaluar primero los conjuntos de datos frente a sus fines originales, ver Figura 14.2, y después considerar lo bien que una combinación de conjuntos de datos que forman el conjunto de datos final se ajusta al concepto y a la población objetivos del resultado estadístico previsto, ver Figura 14.3. Este marco es muy importante en el caso de que se combinen varios

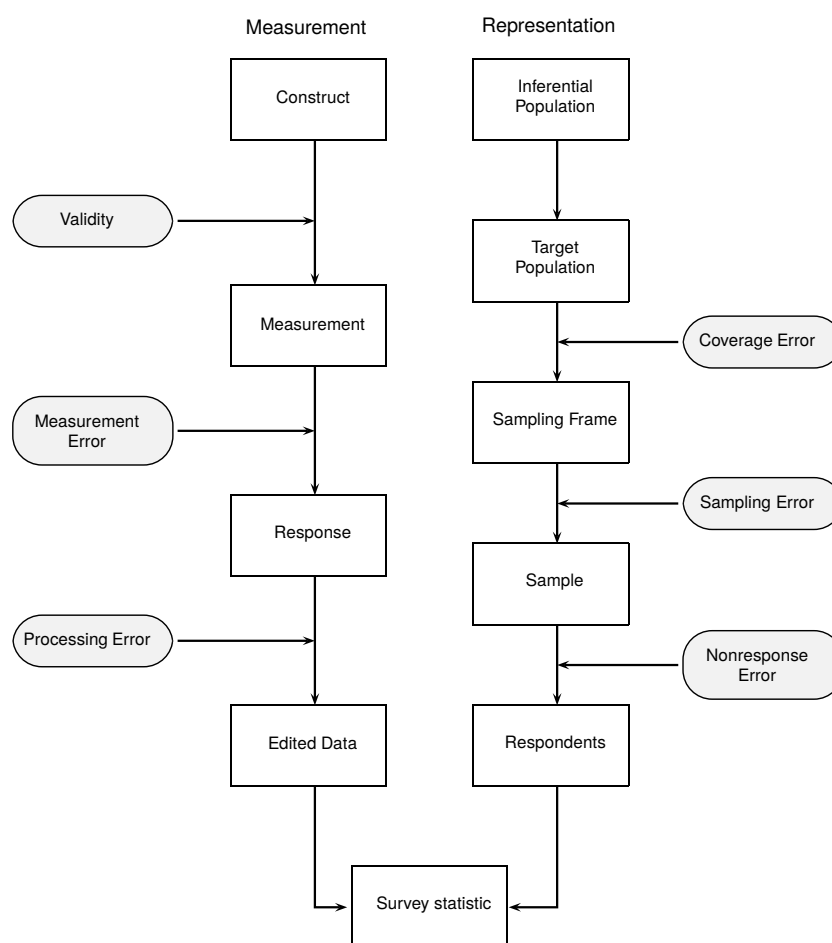


Figura 14.1: Total Survey Error.

conjuntos de datos administrativos o datos administrativos con encuestas para producir una operación estadística. Las variables en los conjuntos de datos administrativos difieren de los datos ideales que nos gustaría usar para medir los conceptos estadísticos objetivos.

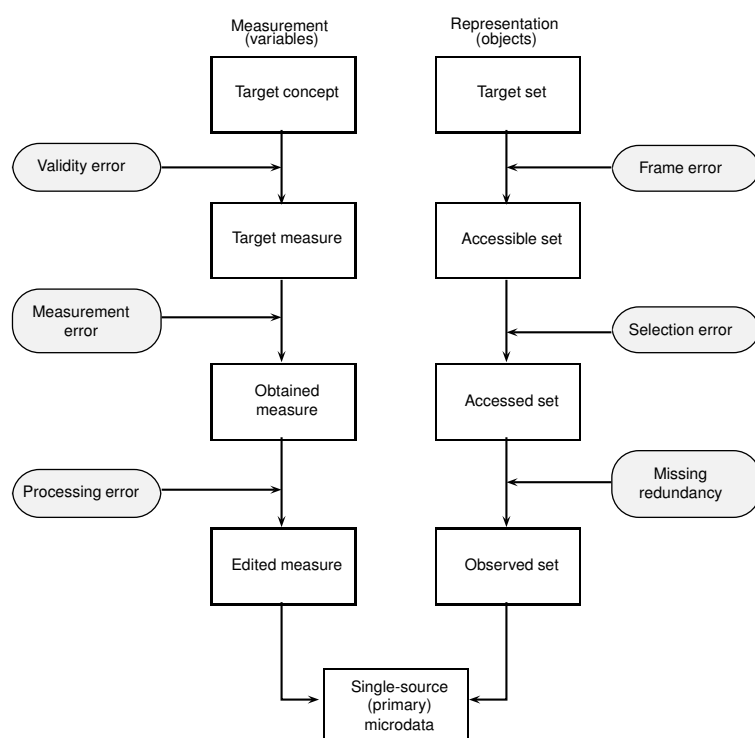


Figura 14.2: Fase 1 del 'ciclo de vida' en dos fases.

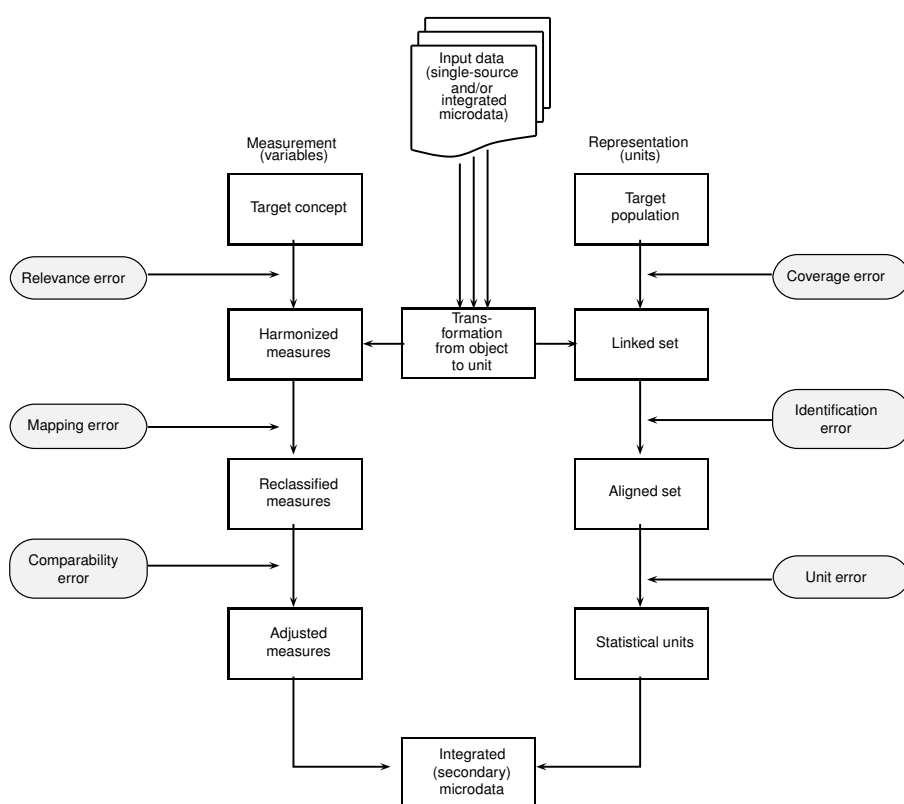


Figura 14.3: Fase 2 del 'ciclo de vida' en dos fases.

## 14.3 Visión conjunta de los métodos

Veamos a continuación algunas ideas sobre el *record linkage*. Aunque las ideas están basadas en modelos estadísticos, el hecho de que los ficheros no estén depurados y la dificultad de desarrollar algunos algoritmos para la estimación y la comparación, han limitado la creación de sistemas de computación que puedan ser usados en una gran variedad de casos.

### 14.3.1 El modelo de *record linkage* de Fellegi-Sunter

Supongamos que partimos de dos ficheros A y B que deseamos cruzar. La idea es clasificar los pares del nuevo espacio  $A \times B$ , formado a partir de los dos ficheros A y B, en  $M$ , que será el conjunto de coincidencias verdaderas, y en  $U$ , el conjunto de no coincidencias verdaderas. Fellegi y Sunter consideraron razones de probabilidades de la forma:

$$R = P(\gamma \in \Gamma | M) / P(\gamma \in \Gamma | U) \quad (14.1)$$

donde  $\gamma$  es un patrón de concordancia arbitrario en el espacio de comparación  $\Gamma$ . Por ejemplo,  $\Gamma$  puede consistir en ocho patrones representando acuerdo o no en la componente más larga del nombre, el nombre de la calle y número de la calle. De forma alternativa, cada  $\gamma \in \Gamma$  también se puede considerar como la frecuencia con la que ocurren valores específicos de componentes del nombre como 'Pérez', 'Cl' o 'Avenida'. El cociente  $R$  se denomina *peso de enlace*, *peso de matching* o *score*.

La regla de decisión viene dada por:

$$\begin{aligned} \text{Si } R > T_\mu, & \text{ entonces se marca el par como una coincidencia.} \\ \text{Si } T_\lambda \leq R \leq T_\mu, & \text{ se marca como una posible coincidencia} \\ & \text{y se señala para una revisión manual.} \\ \text{Si } R < T_\lambda, & \text{ entonces se marca el par como no coincidencia.} \end{aligned} \quad (14.2)$$

Los umbrales  $T_\mu$  y  $T_\lambda$  se fijan en función de unas cotas de error (a priori) de falsas coincidencias y falsas no coincidencias. La regla (14.2) es bastante intuitiva. Si  $\gamma \in \Gamma$  consta básicamente de coincidencias, entonces intuitivamente será más probable que ocurra  $\gamma \in \Gamma$  entre coincidencias que entre no coincidencias y la razón (14.1) será grande. Por contra, si  $\gamma \in \Gamma$  consiste básicamente en discrepancias, entonces la razón (14.1) será pequeña. La regla (14.2) divide el conjunto  $\gamma \in \Gamma$  en tres regiones disjuntas. A la región  $T_\lambda \leq R \leq T_\mu$  la denominaremos región de no decisión o región de revisión manual. En algunas situaciones, los recursos disponibles permiten revisar estos pares de forma manual.

La Tabla 14.1 proporciona ejemplos de pares de registros que deberían de enlazarse usando usando el nombre, la dirección y la edad. Estos pares proporcionan una primera indicación de que los cruces, que deberían de ser sencillos para personal adecuadamente formado, puede no resultar sencillo con las reglas basadas en (14.1) y (14.2). Si el patrón

de concordancia  $\gamma \in \Gamma$  consiste simplemente en que coincidan el nombre, la dirección y la edad, entonces ninguno de los pares coincidirá en ninguno de los tres campos. En la mayoría de las situaciones, una persona suficientemente formada sería capaz de reconocer que los pares pueden ser el mismo pero no sería capaz de cuantificar un *score* para los pares.

Tabla 14.1: Elementos sencillos de coincidencias de pares de registros (dependiendo del contexto)

Nombre	Dirección	Edad
1a. Jose A Perez Garcia	Calle Mayor 16	16
1b. J H Pérez García	C Mayor 16	17
2a. Javier Martínez Moreno	Avenida Principal 49	33
2b. Jabier Martínez Moreno	Avda Ppal 49	36

Si tuviéramos algoritmos de análisis sintáctico<sup>3</sup> programados para separar el campo Nombre en Primer Nombre, Segundo Nombre, Primer Apellido y Segundo Apellido y Dirección en Nombre de la calle, Número de la casa, Piso y Otras componentes, entonces podríamos tener mejores patrones  $\gamma \in \Gamma$  para usar (14.1) y (14.2). Si, además, tuviéramos algoritmos adecuados para comparar campos que contengan errores tipográficos (Javier vs. Jabier), podríamos llegar a un acuerdo parcial con errores tipográficos menores en lugar de decir que la comparación nos conduce a una discrepancia. De forma adicional, son deseables rutinas de estandarización que sustituyan los errores más comunes por una grafía común (C por Calle en el ejemplo 1; Avd, Avenida o Av por Avda en el ejemplo 2). Por último, habría que eliminar los acentos de forma que Perez y Pérez sean coincidentes, lo mismo que Garcia y García, y que Martinez y Martínez.

### 14.3.2 Parámetros de aprendizaje

Los primeros sistemas de *record linkage* que se usaban eran listados administrativos grandes, que debían ser depurados en el sentido de que muchas componentes del nombre, razón social, dirección, y otros campos se revisaban y corregían de manera manual. Con el tiempo, los procesos manuales más sencillos se sustituyeron por procesos informáticos que imitaban a los manuales. Por ejemplo, es muy sencillo convertir apodos en nombres legales ('Paco'  $\rightarrow$  'Francisco') o errores tipográficos obvios ('Péres'  $\rightarrow$  'Pérez').

En prácticamente todas las situaciones reales que podamos conocer, puede no haber disponibilidad de datos de entrenamiento. Los expertos han desarrollado varias formas de obtener parámetros de *record linkage* óptimos sin datos de entrenamiento. En todas menos una de las siguientes secciones, veremos métodos de *aprendizaje no supervisado* en los que no se dispone de datos de entrenamiento.

<sup>3</sup>se traducirá *parsing* por análisis sintáctico



### Las ideas de Newcombe

Las ideas de Newcombe ([Newcombe y Kennedy 1962](#) y [Newcombe, Kennedy y col. 1959](#)) se basan en razones de verosimilitudes. Empezó con un gran registro administrativo que representaba a una población entera, que había sido depurada en el sentido de que los duplicados habían sido borrados y las faltas ortográficas y formatos inconsistentes eliminados. A continuación explicamos la metodología.

Sea el fichero  $C = (c_{ij}), 1 \leq i \leq N_s, 1 \leq j \leq N_c$  un fichero con  $N_s$  registros (filas) y  $N_c$  campos (columnas). Newcombe propone dividir los pares de  $C \times C$  en  $M$  coincidencias y  $U$  no-coincidencias. Aunque conocía la respuesta, quería ser capaz de encontrar la correspondencia entre ficheros externos  $A$  y  $C$  usando las probabilidades condicionadas que calculó al enlazar  $C$  consigo mismo. Si  $A_i$  representa coincidencia en el campo  $i$ ,  $A_i^c$  representa discrepancia en el campo  $i$ , y  $A_i^x$  representa coincidencia o discrepancia, de manera excluyente, en el campo  $i$ . El primer supuesto simplificador de Newcombe es la hipótesis de independencia condicional (IC), que dice que condicionado a estar en el conjunto de coincidencias,  $M$ , o de no coincidencias,  $U$ , la coincidencia en el campo  $i$  es independiente de la coincidencia en el campo  $j$ .

$$(\text{IC}) P(A_i^x \cap A_j^x | D) = P(A_i^x | D) P(A_j^x | D) \quad (14.3)$$

donde  $D$  puede ser tanto  $M$  como  $U$ . Bajo la condición (IC), Newcombe calculó las probabilidades asociadas con cada valor de un campo específico. La intuición es unir los pares de valores comunes en campos individuales. Sea  $(f_{ij}), 1 \leq j \leq I_j$ , las frecuencias específicas (número de valores) del campo  $i$ -ésimo. El número de coincidencias en  $N$  y el de discrepancias es  $N \times N - N$ . Entre las coincidencias  $M$ , hay  $f_{ij}$  pares que coinciden en el  $j$ -ésimo valor del  $i$ -ésimo campo. Entre las discrepancias  $U$ , hay  $f_{ij} \times f_{ij} - f_{ij}$  pares que coinciden en el  $j$ -ésimo valor del  $i$ -ésimo campo. Entonces, la razón de probabilidad de coincidencia en el  $j$ -ésimo valor del  $i$ -ésimo campo es

$$R_{1i} = \frac{P(\text{coincidencia del } j\text{-ésimo valor del } i\text{-ésimo campo} | M)}{P(\text{coincidencia del } j\text{-ésimo valor del } i\text{-ésimo campo} | U)} = \frac{\frac{f_{ij}}{N}}{\frac{f_{ij} \times f_{ij} - f_{ij}}{N \times N - N}} \quad (14.4)$$

Si los pares se toman de dos ficheros,  $A$  y  $B$ , entonces podemos usar  $f_{ij}$  como la frecuencia en  $A$ ,  $g_{ij}$  como la frecuencia en  $B$ , y  $h_{ij}$  como la frecuencia en  $A \cap B$  (que normalmente se aproxima por  $h_{ij} = \min(f_{ij}, g_{ij})$ ), y hacer los cambios correspondientes en (14.4).

### Los métodos de Fellegi y Sunter

[Fellegi y Sunter 1969](#), demostraron la optimalidad de la regla de clasificación dada por (14.2). Su demostración es muy general en el sentido de que se verifica para cualquier representación  $\gamma \in \Gamma$  del conjunto de pares del espacio producto de dos ficheros  $\mathbf{A} \times \mathbf{B}$ . Como observaron, la calidad de los resultados de la regla de clasificación (14.2) depende de la precisión de las estimaciones de  $P(\gamma \in \Gamma | M)$  y  $P(\gamma \in \Gamma | U)$ .

Fellegi y Sunter fueron los primeros en obtener métodos generales para calcular estas probabilidades en situaciones que difieren de las situaciones de Newcombe. Como estos

métodos son muy útiles, describimos lo que establecieron y después mostraremos cómo sus ideas dieron lugar a métodos más generales de *aprendizaje no supervisado* que se pueden usar en un gran número de situaciones.

Fellegi y Sunter observaron varias cosas. En primer lugar,

$$P(A) = P(A|M)P(M) + P(A|U)P(U) \quad (14.5)$$

para cualquier conjunto  $A$  de pares en  $\mathbf{A} \times \mathbf{B}$ . La probabilidad en la izquierda se puede obtener directamente del conjunto de pares. Si los conjuntos  $A$  representan simplemente acuerdo/desacuerdo, bajo la condición (IC), obtenemos

$$P(A_1^x \cap A_2^x \cap A_3^x | D) = P(A_1^x | D)P(A_2^x | D)P(A_3^x | D). \quad (14.6)$$

De esta forma, (14.5) y (14.6) dan lugar a siete ecuaciones con siete incógnitas (ya que  $x$  representa acuerdo o desacuerdo) a resolver. La ecuación (o conjunto de ecuaciones) (14.6) es esencialmente la misma que la ecuación (14.3) y puede ser ampliada a  $K$  campos. Aunque hay ocho patrones asociados con las ecuaciones de la forma (14.6), se puede eliminar una ya que la suma de las probabilidades es uno. En general, con más campos pero aún con coincidencias/discrepancias sencillas entre campos, las ecuaciones se pueden resolver con un algoritmo EM que se verá más adelante en esta sección. Nos referiremos a las probabilidades de la forma  $P(A_i | D)$  como  $m$ -probabilidades si  $D = M$  y como  $u$ -probabilidades si  $D = U$ .

Fellegi y Sunter proporcionaron métodos más generales para emparejamientos basados en la frecuencia que los de Newcombe. De forma específica, obtuvieron las probabilidades generales para acuerdo/desacuerdo y después escalaron las probabilidades basadas en frecuencias a los pesos de acuerdo/desacuerdo. Si  $A_1$  representa acuerdo en el primer campo y  $v_j, 1 \leq j \leq I_1$ , son los valores del primer campo, entonces

$$P(A_1 | D) = \sum_j P(A_1 \cap v_j | D) \quad (14.7)$$

donde  $D$  puede ser tanto  $M$  como  $U$ . Normalmente,  $P(A_i | M) < 1$  para los pesos de acuerdo/desacuerdo en el campo  $i$ . Esto refleja el hecho de que hay menos de un 100 % de acuerdo en el  $i$ -ésimo campo. Aparentemente, podemos pensar en  $1 - P(A_i | M)$  como la tasa media de 'errores tipográficos' en el  $i$ -ésimo campo. Para hacer válida la ecuación (14.7) bajo ciertas restricciones, Fellegi y Sunter asumieron que la tasa de error tipográfico era constante para todos los valores  $v_j, 1 \leq j \leq I_1$ , asociados con el  $i$ -ésimo campo. Las ideas basadas en la frecuencia de Fellegi y Sunter fueron ampliadas por Winkler (W. E. Winkler 1990) con cálculos en caso de hipótesis más débiles.

Hay un número de hipótesis implícitas que se hacen a menudo en el caso de cruce de dos ficheros y de cálculo de probabilidades usando (14.5) - (14.7). La primera es que hay un gran solapamiento entre los dos ficheros  $A$  y  $B$ . Esto básicamente significa que  $A \cup B$  no coincide ni mayormente con  $A$  ni mayormente con  $B$ . Si esta hipótesis no es

cierta, entonces las probabilidades obtenidas mediante los métodos de Newcombe y los de Fellegi y Sunter pueden no funcionar bien. La segunda hipótesis es que ningún fichero  $A$  ni  $B$  pueden ser de forma simultánea muestras de dos ficheros más grandes  $A_2$  y  $B_2$ . La tercera hipótesis es que las tasas de errores tipográficos son bastante bajas de forma que sean válidos los cálculos basados en la frecuencia, que se basan en que los distintos valores observados de los campos son válidos. Si un valor relativamente raro en un apellido tiene distintas formas de escribirse, entonces no es posible calcular de forma correcta las probabilidades basadas en la frecuencia directamente del fichero.

En la práctica, es necesario realizar *bloqueos* de dos ficheros para integrar los pares. Si dos ficheros  $A$  y  $B$  contienen cada uno 10.000 registros, entonces hay  $10^8$  pares en el producto  $A \times B$ . Hasta hace poco no se podían llevar a cabo cálculos de  $10^8$  pares. Realizar *bloqueos* significa que sólo consideramos los pares en los que coinciden determinadas características. Por ejemplo, podemos considerar únicamente los pares en los que coinciden la primera letra del primer nombre, el último nombre (si hay más de un nombre), y la fecha de nacimiento, o la razón social, la dirección y el teléfono. Si creemos (posiblemente basándonos en experiencias anteriores) que no estamos consiguiendo una proporción suficientemente grande de coincidencias con este primer criterio de bloqueo, podemos intentarlo con un segundo. Por ejemplo, podemos considerar parejas en las que coincidan la primera letra del primer nombre, la primera letra del último nombre (si hay más de un nombre), y el código postal a 5 dígitos, o la razón social, la provincia y el código a tres dígitos del municipio.

### El algoritmo EM

A continuación se explicará, aunque sin entrar en mucho detalle, el algoritmo básico de EM aplicado a *record linkage*, describiendo sus limitaciones y algunos de sus alcances.

Para cada  $\gamma \in \Gamma$ , consideramos

$$\begin{aligned} P(\gamma) &= P(\gamma|M)P(M) + P(\gamma|U)P(U) \\ P(\gamma) &= P(\gamma|C_1)P(C_1) + P(\gamma|C_2)P(C_2) \\ P(\gamma) &= P(\gamma|C_1)P(C_1) + P(\gamma|C_2)P(C_2) + P(\gamma|C_3)P(C_3) \end{aligned}$$

y observamos que la proporción de pares representados por  $\gamma \in \Gamma$  (es decir, el lado izquierdo de la ecuación (14.7) se puede calcular directamente a partir de los datos disponibles. En cada variante, tanto  $M$  y  $U$ ,  $C_1$  y  $C_2$ , o  $C_1$ ,  $C_2$  y  $C_3$  son particiones de  $A \times B$ .

Si el número de campos asociados con  $P(\gamma)$  es  $K > 3$ , entonces podemos resolver la combinación de ecuaciones dadas por (14.7) y (14.6) usando el algoritmo EM. Aunque hay métodos alternativos para resolver las ecuaciones, como los métodos de los momentos y de mínimos cuadrados, el EM es preferible por su estabilidad numérica. Bajo IC, la programación se simplifica y el cálculo se reduce considerablemente (de  $2^k$  a  $2k$ ).

Debemos tener cuidado al aplicar el algoritmo EM a datos reales. El algoritmo EM que se ha aplicado al *record linkage* es un *algoritmo de clases latentes* que busca dividir  $A \times B$  en los conjuntos de pares buscados  $M$  y  $U$ . La probabilidad de un indicador de clase que determine si un par está en  $M$  o en  $U$  es el dato que debe ser estimado junto con las  $m$ -probabilidades y las  $u$ -probabilidades. Puede ser necesario aplicar el algoritmo EM a un subconjunto particular  $S$  de pares de  $A \times B$  en el que la mayoría de las coincidencias  $M$  estén concentradas, en el que los campos utilizados para el cruce se puedan separar  $M$  de  $U$  de forma clara, y para el cual se puedan elegir probabilidades iniciales adecuadas. Puesto que el EM es un algoritmo de maximización local, puede ser necesario que las probabilidades iniciales se tengan que elegir con cuidado en base a experiencias con ficheros similares. Al ser el algoritmo EM de clases latentes un algoritmo general de conglomerados, no hay seguridad de que el algoritmo divida  $A \times B$  en dos clases  $C_1$  y  $C_2$  que coincidan casi exactamente con  $M$  y  $U$ .

### 14.3.3 Comparadores de cadenas

En la mayoría de las situaciones de *matching*, conseguiremos malos resultados si comparamos dos cadenas de manera exacta (carácter a carácter, es decir, letra a letra o número a número) debido a los errores tipográficos. Abordar los errores tipográficos usando de manera aproximada la comparación de cadenas ha sido un proyecto muy importante desde el punto de vista informático (véase, por ejemplo, [Hall y Dowling 1980](#) y [Navarro 2001](#)). En el *record linkage* necesitamos una función que represente acuerdo aproximado, representando el acuerdo por 1 y los distintos grados de acuerdo parcial por números entre 0 y 1. También necesitamos ajustar la razón de verosimilitud (14.1) de acuerdo con los valores de acuerdo parcial. Disponer de estos métodos es crucial para llevar a cabo el proceso de emparejamiento o *matching*.

[Jaro 1989](#) introdujo un comparador de cadenas que tiene en cuenta las inserciones, las eliminaciones y las transposiciones. El algoritmo básico de Jaro tiene tres componentes: 1. calcula la longitud de cadenas, 2. encuentra el número de caracteres comunes en las dos cadenas, y 3. encuentra el número de transposiciones. La definición de común es que el número de caracteres coincidentes debe de ser la mitad de la longitud de la cadena más corta. La definición de transposición es que el carácter de una cadena está desordenado con respecto al correspondiente carácter común de la otra cadena. El valor del comparador de cadenas (escalado por consistencia con la práctica en términos computacionales) es:

$$\Phi_j(s_1, s_2) = \frac{1}{3} \left( \frac{N_C}{len_{s_1}} + \frac{N_C}{len_{s_2}} + 0,5 \frac{N_t}{N_C} \right), \quad (14.8)$$

donde  $s_1$  y  $s_2$  son las cadenas de longitudes  $len_{s_1}$  y  $len_{s_2}$ , respectivamente,  $N_C$  es el número de caracteres comunes entre  $s_1$  y  $s_2$  donde la distancia para comunes es la mitad de la distancia mínima entre  $s_1$  y  $s_2$ , y  $N_t$  es el número de trasposiciones.

Usando conjuntos de datos reales, [W. E. Winkler 1990](#) introdujo métodos para modelizar cómo los distintos valores del comparador de cadenas afecta a la verosimilitud ([14.1](#)) en la regla de decisión de Fellegi-Sunter. Winkler también demostró cómo una variante del comparador de cadenas de Jaro  $\Phi$  mejora la eficacia del emparejamiento en comparación con aquellas situaciones en las que los comparadores de cadenas no se utilizan.

### 14.3.4 Datos de entrenamiento

Raramente se dispone de un conjunto de datos de entrenamiento representativos a la hora de obtener los parámetros para las reglas de clasificación. Si un conjunto de datos está disponible, entonces es posible conseguir los parámetros añadiendo cantidades apropiadas para obtener las probabilidades en ([14.1](#)) y ([14.2](#)). De hecho, con suficientes datos de entrenamiento, es posible estimar directamente las probabilidades en ([14.1](#)) que explican las dependencias entre los distintos campos de cruce y estimar las tasas de error.

## 14.4 Preparación de los datos

En los proyectos de *matching*, preparar los datos de los dos ficheros A y B en un formato consistente de forma que los datos se puedan usar con un software de *record linkage* a menudo requiere más trabajo (de 3 a 12 meses) que las operaciones propias de *matching* (de 1 a 3 semanas). La falta de formación, de tiempo o de recursos para mejorar los ficheros son a menudo las razones por las que los proyectos de *matching* fracasan. Veremos, a continuación, algunos aspectos importantes sobre la obtención, preparación y estandarización de ficheros.

### 14.4.1 Descripción de un proyecto de *matching*

Construir un marco o un registro administrativo para un país o una región grande implica muchos pasos. Lo mismo ocurre con los métodos de construcción, que precisan pares de listas, o el localizar duplicados en una lista dada.

1. Identificar las listas existentes que se pueden usar para crear la lista principal. En este caso es importante concentrarse en 10 listas o menos. Es prácticamente inviable considerar cientos de listas.
2. Para cada lista es necesario tener un diseño de registro. La información que debería incluir es la posición de las variables así como los posibles valores que pueden tomar. Por ejemplo, una lista puede tener varios códigos de estado indicando si la entidad está activa (para el caso de unidades económicas como empresas) o viva (para el caso de personas). En el caso de un listado de empresas, puede existir un código de estado adicional que indique si el registro está asociado con otra entidad como subsidiario o duplicado (por ejemplo, si tenemos un listado de empresas y otro listado de establecimientos, algunas empresas pueden tener

un único establecimiento, pero otras pueden estar multilocalizadas). Si el diseño de registro no está disponible, entonces es mejor desechar la lista. Si la lista no es compatible con el sistema informático o tiene un formato incompatible, entonces es mejor desechar la lista.

3. Obtener las listas para ponerlas en un formato compatible que se usará en los programas de detección de duplicados y de actualización. Si el listado no verifica los programas de estandarización de nombres y direcciones, entonces se debería de considerar el rechazar el listado o usar sólo aquellos registros que puedan ser estandarizados. El formato estándar debería incluir un campo para la fuente de la lista y la fecha de la misma. Si es posible, es una buena idea tener una fecha para los registros individuales de la lista.
4. Si los recursos lo permiten, se puede obtener mayor precisión enlazando cada fuente potencial de actualización a la lista principal secuencialmente. Enlazar cada lista de manera secuencial permite una limpieza manual de duplicados más acurada. Si la limpieza manual no se puede llevar a cabo de manera eficiente, entonces los duplicados en la lista principal producirán más y más duplicados a medida que la lista principal se va actualizando. Si parece que una lista individual está dando lugar a que demasiados duplicados sean erróneamente incluidos en la lista principal, entonces es mejor rechazar el listado como una fuente para actualizar. Si un subconjunto grande, obtenido a partir de una fuente actualizada, no da lugar a un número suficientemente grande de nuevas entidades en la lista principal, entonces también sería mejor rechazarlo.
5. Después de un *matching* inicial, se deberían realizar de manera sistemática procesos adicionales informatizados y manuales para seguir identificando duplicados en la lista principal. Un proceso muy útil es asegurarse de que la representación de nombres y direcciones asociados con un organismo están escritos en su forma más útil y sin errores tipográficos. Estos procesos de mejora adicionales se deberían usar de manera continua. Si las actualizaciones y limpiezas de listas con muchas empresas pequeñas se realiza de manera manual, entonces la calidad general de la lista se puede deteriorar de forma añadida en cada actualización debido a que las empresas pequeñas suelen cambiar de nombre o dirección con mayor frecuencia.

Muchos proyectos de *matching* fracasan porque ni siquiera se consigue pasar del primer o segundo pasos que se acaban de exponer. Mantener los listados actualizados puede resultar muy difícil. En EEUU se estima que en los listados de pequeñas empresas, el cambio de nombre o de dirección puede superar el 10 % al año.

#### 14.4.2 Preparación inicial de ficheros

Para obtener los ficheros, el primer problema es determinar si los ficheros están en ficheros secuenciales (formato plano), en una base de datos o en ficheros SAS. Muchos

programas de *record linkage* están diseñados sólo para ficheros secuenciales, y los ficheros en otros formatos tendrán que ser transformados. Si se trabaja con muchos ficheros es necesario tener un formato estándar y procedimientos para que los ficheros estén en el formato más compatible para el *record linkage*. El diseño de registro nos dará las descripciones de los campos individuales que se pueden comparar. Por ejemplo, el código para el sexo puede ser del tipo Sexo1 (hombre=H, mujer=M, missing=b) o Sexo2 (hombre=1, mujer=2, missing=0). A través de programas sencillos se puede pasar de un tipo de codificación a otra.

En los ficheros con un buen mantenimiento suele haber un código de estado indicando si una unidad está aún viva (para el caso de personas físicas) o activa (para el caso de unidades jurídicas) y si la información (como el teléfono o la dirección) está actualizada. Si un fichero tiene un código de estado indicando si un registro está fuera de ámbito, entonces en la mayoría de las aplicaciones de *matching* habrá que eliminar los registros fuera de ámbito antes de usar el fichero para actualizar o hacer una combinación. Por ejemplo, las compañías eléctricas tienen muy buena información sobre las direcciones de las viviendas. Es posible que esa lista contenga también información sobre pequeños comercios, por tener el mismo tipo de tarifas. Si se quiere usar la lista de viviendas de las compañías eléctricas podemos encontrarnos con muchas direcciones de comercios que estarían fuera de ámbito en el caso de marcos de encuestas dirigidas a individuos/hogares.

Puede ser necesario revisar varios campos entre dos ficheros. Por ejemplo, si un fichero tiene direcciones que están compuestas de un nombre de calle y un número y otro fichero tiene direcciones que son apartados de correos, entonces puede ser difícil cruzar los dos ficheros usando la información del nombre y la dirección postal. Con listados de empresas puede ser necesario tener información auxiliar que permita distinguir la sede central del resto de sedes, como por ejemplo las fábricas. Dependiendo del tipo de unidad a la que se dirige el cuestionario puede ser necesario un listado de establecimientos o sólo la sede central. Y en ocasiones si se manda el cuestionario tanto a la sede central como al resto de establecimientos se podrían estar duplicando los datos.

Veamos un conjunto de procedimientos que se pueden usar para la preparación preliminar de ficheros y que se pueden realizar directamente con programas informáticos. Estos procedimientos de depuración y comprobación de consistencia de los ficheros se conoce como *estandarización* y consta de los siguiente pasos.

1. La sustitución de variantes ortográficas por una ortografía consistente se denomina *estandarización ortográfica*.
  - Sustituir 'Doctor' o 'Dr.' por 'Dr'.
  - Sustituir 'Paco' o 'Curro' por 'Francisco'.
  - Sustituir 'Avenida' o 'Avda.' por 'Avda'.



2. La sustitución de códigos inconsistentes se denomina asegurar la *consistencia de códigos*.

- Sustituir Sexo (hombre=1, mujer=2, missing=0) por Sexo (hombre=H, mujer=M, missing=' ').
- Sustituir '11 Enero 1999' o 'Enero 11 1999' por MMDDYYYY='01111999' o YYYYMMDD='19990111'.

La consistencia de códigos es la consistencia del estado de valores de las variables (o campos). En el *record linkage*, una variable (o campo) generalmente es una cadena de caracteres, tales como un nombre completo, una dirección completa o un componente como el nombre o los apellidos.

3. La identificación la posición inicial y final de los componentes individuales de una cadena de forma libre tales como un nombre o una dirección se la denomina *parsing*.

- Identificar la posición del nombre, el primer apellido y el segundo apellido en 'Sr José María García Fernández' y 'José María García Fernández'.
- Identificar la posición de la calle y el número en 'Paseo de la Castellana 183' o 'Paseo de la Castellana 183, 636'.

La idea del análisis sintáctico es permitir las comparaciones de campos (variables) que deberían ser consistentes y fáciles de comparar. No es fácil comparar nombres y direcciones en formato libre excepto de forma manual. Las tres ideas anteriores de estandarización son a menudo previas a situaciones en que los nombres y direcciones en formato libre se descomponen (analizan sintácticamente) en sus componentes. Veamos en los siguientes apartados la estandarización general de nombres y direcciones.

### 14.4.3 Estandarización y análisis sintáctico de nombres

La estandarización consiste en sustituir varias formas de escribir palabras por una única. Por ejemplo, distintas formas de escribir 'Compañía' se pueden sustituir por una única forma estandarizada 'Cía.' Los componentes de software de estandarización pueden separar una cadena como un nombre o una dirección en palabras (es decir, conjuntos de caracteres que están separados por espacios u otros delimitadores). Cada palabra es entonces comparada en tablas de referencias para obtener una forma de escribirla estándar. La parte superior de la Tabla 14.2 muestra varios nombres comunes que se reemplazan por grafías estandarizadas. Después de la estandarización, la cadena del nombre se analiza sintácticamente en sus componentes (segunda mitad de la tabla) que puede ser ya comparada.

Existen nuevos métodos que usan modelos de cadenas de Markov ocultas, lo que permite mejorar la estandarización de los nombres basadas en las reglas.



Estandarizado								
1	DR John J Smith MD							
2	Smith DRY FRM							
3	Smith&Son ENTP							
Tras el análisis sintáctico								
	Pre	Nombre	Mid	Apellido	Post1	Post2	Bus1	Bus2
1	DR	John	J	Smith	MD			
2				Smith			DRY	FRM
3				Smith		Son	ENTP	

Tabla 14.2: Ejemplos de análisis sintáctico de nombres

#### 14.4.4 Estandarización y análisis sintáctico de direcciones

La Tabla 14.3 muestra una posible estandarización de direcciones. La primera mitad de la tabla muestra unas cuantas direcciones que han sido estandarizadas. En la estandarización, palabras comunes como 'Calle' se sustituyen por una abreviatura adecuada como 'Cl' que puede considerarse como una grafía estándar que puede explicar algunos errores de ortografía. La segunda mitad de la tabla representa los componentes de las direcciones producidas por el análisis sintáctico.

Estandarizado										
1	16 W Main ST APT 16									
2	RR 2 BX 215									
3	Fuller Bldg Suite 405									
4	14588 HWY 16 W									
Tras el análisis sintáctico										
	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
1	W	16	Main			ST		16		
2				2	215					
3									405	Fuller
4		14588	HWY	16			W			

Tabla 14.3: Ejemplos de análisis sintáctico de direcciones

#### 14.4.5 Estandarización y normalización de registros administrativos

Para combinar registros administrativos con datos estadísticos basta con que exista un único identificador para cada entidad (persona, empresa, viviendas) que se pueda usar como variable de enlace <sup>4</sup> entre los registros administrativos y las bases de datos estadísticas. Desgraciadamente, en la práctica esta idea presenta muchas dificultades,

<sup>4</sup>se traduce *linking variable* como variable de enlace

ya que no siempre este identificador estará disponible, por ejemplo el D.N.I. no es obligatorio para los menores de 14 años, y en muchos casos no está correctamente estandarizado, por ejemplo en el caso de los pasaportes, véase [Teijeiro Alfonsín 2007](#).

A continuación se detallan algunas fases esenciales que permiten realizar de forma satisfactoria la estandarización y la normalización de registros administrativos.

### Definición de un orden numérico

Antes de llevar a cabo ninguna otra tarea, es conveniente asignar un número de orden (NORDEN) a cada una de las componentes (filas u observaciones) del fichero a tratar. Este número de orden actuará como identificador único de cada observación del registro y facilitará la detección y el posterior tratamiento de incidencias, especialmente de duplicidades. Una buena práctica será construir un número de orden mediante la asignación de códigos numéricos consecutivos e independientes. Se puede usar alguna variable numérica para esta ordenación, por ejemplo el código para cada provincia o región.

En la Tabla 14.4 se puede ver un ejemplo en el que se usa como variable numérica el código de provincia, que nos permite seguir un orden consecutivo en la codificación. Los metadatos juegan un papel muy importante ya que el uso de clasificaciones oficiales proporciona información pública y perfectamente definida. En este caso particular, se usa la clasificación de las provincias españolas <sup>5</sup>.

Antes de la asignación		Tras la asignación		
Identificador	Provincia	NORDEN	Identificador	Provincia
N0X9011655-R	01	0100000001	N0X9011655-R	01
P0X-014697	01	0100000002	P0X-014697	01
X000241/86	02	0200000001	X000241/86	02
X000327/86	02	0200000002	X000327/86	02
521382901	03	0300000001	521382901	03

Tabla 14.4: Ejemplos de asignación de números de orden

### Estandarización del identificador

Los códigos que se utilizan oficialmente para identificar a personas o empresas, normalmente están formados por una cadena alfanumérica con una determinada longitud y formato. Algunos como el Número de la Seguridad Social de las personas físicas o el Código de Cuenta de Cotización a la Seguridad Social de las empresas están formados sólo por dígitos, mientras que en el caso de D.N.I. o el C.I.F. se trata de un código numérico junto con algunas letras, obtenidos a partir de algoritmos. Por último, puede tratarse de cadenas alfanuméricas pero que no siguen ningún estándar, como ocurre en

<sup>5</sup>[https://ine.es/daco/daco42/codmun/cod\\_provincia.htm](https://ine.es/daco/daco42/codmun/cod_provincia.htm)

el caso de los Pasaportes o del N.I.E.

En algunos registros el código identificador puede tener errores y, por tanto, será necesaria una transformación antes de su uso. Si esta fase no se lleva a cabo de forma cuidadosa, todo el trabajo posterior de cruce de ficheros resultará afectada ya que habrá registros que figurarán en varios ficheros, pero no será posible establecer la conexión si la codificación es distinta en cada fichero. También podría pasar que el registro con el identificador erróneo fuese considerado un nuevo elemento y se incorporase como un alta; por ejemplo, en el caso de que haya ceros a la izquierda y se confundan estos se confundidos con la vocal mayúscula O.

Hay casos en los que es preferible tener varios candidatos, incluso si no todos ellos son válidos, a perder al correcto. Por ejemplo, cuando se trabaje con registros con variables de identificación incompletas, registros con identificadores no disponibles o identificadores no óptimos numéricamente o incluso en el caso de registros con identificadores numéricos correctos pero en los que podría ser preferible enlazar usando otra variable de enlace distinta. En todos estos casos es recomendable considerar múltiples candidatos y después usar un algoritmo de selección para obtener el correcto. Por ejemplo, podemos disponer de varios ficheros con información sobre establecimientos industriales, que contienen el N.I.F. de cada establecimiento, así como las variables de identificación (razón social, dirección postal, municipio, provincia, etc.). En primer lugar se puede hacer un primer cruce eliminando en el N.I.F. la primera letra, ya que podría darse el caso de que alguna empresa hubiese pasado de sociedad limitada a sociedad anónima, o viceversa, y que algún fichero no esté actualizado. Posteriormente, se podría realizar un segundo cruce por provincia y municipio, dirección postal, etc.

Consideramos pues, que si un individuo en cuestión existe en varios ficheros la conexión será posible. Un proceso de estandarización de cualquier cadena alfanumérica que verifica el esquema de trabajo mencionado puede consistir en los siguientes pasos:

1. En primer lugar, cada letra se transforma en mayúscula/minúscula, los caracteres raros se eliminan y la cadena total se separa en tantas partes como cambios se hayan observado de bloques numéricos a alfanuméricos, o viceversa. Véase [Amón Uribe y Jiménez 2010](#) y [A. K. Elmagarmid 2007](#).
2. En segundo lugar, se pueden asignar los identificadores a cada individuo. Se recomienda considerar tantos identificadores como combinaciones sean posibles a partir de los bloques numéricos obtenidos en el paso anterior.
3. Si algún divisor alfanumérico entre dos bloques numéricos es una 'O', uno de los códigos numéricos resultantes obtenidos será el resultado de reemplazar esa vocal mayúscula 'O' por el número '0'.

En el ejemplo de la Tabla 14.5 cada observación es identificada de manera unívoca con el número de orden (NORDEN) pero con más de un código numérico estandarizado. Estos códigos numéricos se corresponden con la parte numérica del identificador oficial y se usarán para unirlos con otros registros en los que figuran también identificadores

NORDEN	IDEN	Bloques numéricos		Identificadores numéricos estandarizados			
		NUM1	NUM2	COD1	COD2	COD3	COD4
0100000001	N0X9011655-R	0	9011655	9011655	-	-	-
0200000001	X000241/86	241	86	241	86	24186	-
0200000002	X000327-86	327	86	327	86	32786	-
0400000001	D07816O423	07816	423	07816	423	7816423	78160423

Tabla 14.5: Ejemplo de extracción de códigos numéricos estandarizados

estandarizados.

### Estandarización del nombre

Este procedimiento se puede aplicar a cualquier cadena alfabética, es decir, nombre, apellidos, razones sociales de empresas, direcciones postales, etc. El hecho de disponer de nombres estandarizados resulta imprescindible en el uso de registros administrativos. El simple hecho de incorporar por error un espacio en blanco al principio de un nombre es un obstáculo en el *record linkage*. Mientras que para el ojo humano la identificación de un espacio al principio de un nombre es algo directo, cualquier ordenador considerará como dos nombre distintos el que tiene y el que no tiene el espacio al principio.

A pesar de que este tipo de inconvenientes impide muchas comparaciones, la aplicación de una serie de reglas muy sencillas puede mejorar la comparabilidad entre cadenas alfabéticas. La regla más sencilla es la simple conversión a mayúsculas/minúsculas de todas las letras. Además, es conveniente:

- Analizar el contenido de las cadenas convirtiendo símbolos extraños, que de forma sistemática se confunden con ciertas letras. Por ejemplo, es posible que la vocal mayúscula 'O' se sustituya por el número '0'.
- Limpiar las cadenas, eliminando los acentos y las diéresis (ÁÉÍÓÚ, ÀÈÌÒÙ, ÄËÏÖÜ, ÂÊÎÔÛ), y eliminando también los símbolos no alfabéticos (. · + - / : ; # @ = ) así como espacios innecesarios, como dobles espacios. Véase [Amón Uribe y Jiménez 2010](#) y [A. K. Elmagarmid 2007](#).
- Extraer las partes que aparecen en las cadenas pero que no son nombres por ellos mismos (artículos, preposiciones, contracciones). Para direcciones postales es importante eliminar de sus nombres aquellas partes que se refieren al tipo de vía (Calle, Plaza, Avenida) y almacenarlas en un campo independiente para posibles uniones con otros ficheros. De igual forma, en el caso de las razones sociales de empresas es pertinente separar los acrónimos referidos al tipo de empresa (SA, S.A., SL, S.L.).
- Disponer de diccionarios que permitan transformaciones o codificaciones de abreviaciones, iniciales, apodos, diminutivos, nombres compuestos y cualquier variación léxica posible o composición de los nombres considerados. La situación ideal es aquella en la que ya existe un diccionario establecido con directorios

o censos susceptibles de ser usados como referencia. Pero si no lo tenemos, la construcción de uno básico que tenga en cuenta las equivalencias más frecuentes servirá para estandarizar un gran volumen de elementos. Con un pequeño esfuerzo se mejorará la comparabilidad de campos en distintos ficheros. Ejemplos típicos de estas situaciones son abreviaciones del nombre 'María' o tipos de vías como 'Calle', 'Plaza' o 'Avenida'.

#### 14.4.6 Resumen sobre el preprocesamiento

En ocasiones puede resultar complicado el preprocesamiento de los datos de algunos ficheros. En situaciones en las que los datos tienen una alta calidad, el preprocesamiento puede dar lugar a una gran mejora en cuanto a la eficacia del *matching*, mayor que el uso de comparadores o de parámetros 'optimizados'. Se ha llegado a obtener hasta un 90 % de mejora en la eficacia del *matching* por el preprocesamiento.

### 14.5 Caso práctico con registros administrativos

Una vez que las principales variables de identificación han sido estandarizadas, el registro está listo para llevar a cabo el proceso de *record linkage*. Si los ficheros a unir tienen identificadores de buena calidad, su mera unión a través de este identificador puede ser suficiente para combinarlos. Pero en ocasiones esto no ocurre (puede que no se disponga del D.N.I. o que el N.I.F. no esté actualizado) por lo que se puede recurrir a lo siguiente:

- Seleccionar varias variables y usar combinaciones de ellas para cruzar los ficheros. Por ejemplo, en el caso de las personas se pueden usar el nombre, el primer apellido, el segundo apellido y la dirección, y en el caso de unidades económicas se puede usar la dirección postal, el municipio y la actividad económica.
- Considerar para cada registro individual del primer fichero todos los posibles candidatos en el segundo fichero y a continuación seleccionar el óptimo.

Si tenemos a nuestra disposición varios tipos de identificadores, el *matching* puede ser realizado en varios pasos, es decir, realizando el cruce en una primera fase con el identificador más frecuente o fiable, y en fases sucesivas con el resto de identificadores ordenados en orden descendente según la fiabilidad.

Si se dispone de más de un campo de identificación adecuado, existen dos posibilidades de linking. Una de ellas es realizar varias uniones, que pueden ser ejecutadas a través de esos diversos nombres disponibles. La otra es seleccionar sólo uno o más campos como enlace para unir la información y después usar el resto de campos como información auxiliar para que el algoritmo seleccione el candidato óptimo. La opción más apropiada debe estar basada en la situación de los ficheros y el objetivo de su unión. En el caso de que existan varios candidatos será necesario seleccionar un algoritmo adecuado que permita elegir al candidato ideal. Para este propósito son necesarias funciones de distancia o similaridad adecuadas. Las funciones de distancia o similaridad deben

proporcionar un valor numérico que indique el grado de similitud entre dos cadenas alfanuméricas (Amón Uribe y Jiménez 2010 y A. K. Elmagarmid 2007). Existen múltiples opciones como, por ejemplo, la distancia Levenshtein (V. I. Levenshtein 1965) o la de Jaro-Winkler (W. E. Winkler 1990). La elección de la función de distancia se basa en el tipo de variable a comparar. En el caso de disponer de varios identificadores, se recomienda tener una función de distancia adecuada para cada campo de identificación que se vaya a usar para el *record linkage*: una para evaluar cadenas alfanuméricas, como el nombre, la razón social, el D.N.I., el N.I.F., y otra para evaluar fechas, otra para evaluar códigos de municipios, etc. Una vez que se dispone del valor numérico que indica la similitud o divergencia, el algoritmo de selección para el candidato óptimo se define en función de unas reglas de decisión que dependerán de las combinaciones de los *scores* obtenidos para cada uno de los campos comparados.

## Bibliografía

- A. Fresneda Pacheco, M. Pérez Julián (2014). *Use of administrative records in official statistics: a practical view*. URL: <http://www.seio.es/BEIO/Use-of-administrative-records-in-official-statistics-a-practical-view-2.html>.
- A. K. Elmagarmid P. G. Ipeirotis, V. S. Verykios (2007). "Duplicate Record Detection: A Survey". En: *IEEE Transactions on Knowledge and Data Engineering* 19.1, págs. 1-16. DOI: 10.1109/TKDE.2007.250581.
- Amón Uribe, I. y C. Jiménez (2010). *Detección de duplicados: una guía metodológica*. URL: <https://revistas.unab.edu.co/index.php/rcc/article/view/1387>.
- Fellegi, I.P. y A.B. Sunter (1969). "A theory for record linkage". En: *Journal of the American Statistical Association* 64, págs. 1183-1210.
- Hall, P.A.V. y G.R. Dowling (1980). "Approximate string comparison. Association of Computing Machinery". En: *Computing Surveys* 12, págs. 381-402.
- Jaro, M.A. (1989). "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida". En: *Journal of the American Statistical Association* 89, págs. 414-420.
- Navarro, G. (2001). "A guided tour of approximate string matching". En: *Association of Computing Machinery Computing Survey* 33, págs. 31-88.
- Newcombe, H.B. y J.M. Kennedy (1962). "Record linkage: making maximum use of the discriminating power of identifying information". En: *Communications of the Association for Computing Machinery* 5, págs. 563-567.
- Newcombe, H.B., J.M. Kennedy, S.J. Axford y A.P. James (1959). "Automatic linkage of vital records". En: *Science* 130, págs. 954-959.
- Statistics New Zealand (2016). *Guide to reporting on administrative data quality*. ISBN: 978-0-908350-29-2. URL: <https://www.stats.govt.nz/methods/guide-to-reporting-on-administrative-data-quality>.
- Teijeiro Alfonsín, E. (2007). *Sobre el aprovechamiento de registros administrativos en la realización de Censos de Población y Vivienda*. URL: <https://dialnet.unirioja.es/servlet/articulo?codigo=2739910>.
- V. I. Levenshtein (1965). "Binary codes capable of correcting deletions, insertions, and reversals". En: *Dokl. Akad. Nauk SSSR* 163, págs. 845-848.

- W. E. Winkler (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". En: *Proceedings of the Section on Survey Research Methods*, págs. 354-359.
- Wallgren, A. y B. Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.
- Winkler, W.E. (2006). *Overview of record linkage and current research directions*. URL: <http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf>.
- (2009). *Record linkage*, en D. Pfefferman and C.R. Rao (eds.), *Handbook of Statistics 29A*, cap. 14, pp. 351-370. Amsterdam: North-Holland.

## Tema 15

### Metadatos de la producción estadística. I. GSBPM. Introducción. El modelo. Relaciones con otros modelos y estándares. Niveles 1 y 2 del GSBPM. Descripciones de fases y subprocesos (fases 1 a 3).

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

UNECE (2019c). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>

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.

#### 15.1 Introducción

El *Generic Statistical Business Process Model* (GSBPM) (UNECE 2019c) es un modelo de procesos de negocio<sup>1</sup> que describe y define de manera genérica el conjunto de tareas de producción necesarias para elaborar estadísticas oficiales. Proporciona un marco estándar y terminología armonizada para ayudar a las organizaciones estadísticas a modernizar sus procesos de producción estadísticos, así como a compartir métodos y componentes del proceso.

El GSBPM también se puede usar para integrar estándares de datos y de metadatos, como una plantilla para documentar el proceso, para armonizar infraestructuras informáticas estadísticas y para proporcionar un marco para la evaluación y mejora de la calidad del proceso.

---

<sup>1</sup>Traducimos *business process* como proceso de negocio; también se encuentra a veces como proceso de trabajo, proceso operativo o proceso de empresa, con ligeras variantes.





El GSBPM se debe, por tanto, concebir más como una matriz, a través de la cual hay muchas posibles rutas. En este sentido, el GSBPM aspira a ser lo suficientemente genérico como para ser ampliamente aplicable a muchos tipos de operaciones estadísticas, promoviendo así un punto de vista estandarizado del proceso estadístico en los institutos de estadística, sin llegar a ser ni demasiado restrictivo ni demasiado abstracto y teórico.

### 15.2.1 La estructura

El GSBPM abarca tres niveles:

- Nivel 0, el proceso estadístico;
- Nivel 1, las ocho fases del proceso estadístico;
- Nivel 2, los sub-procesos de cada fase.

Los niveles 1 y 2 están representados en la Figura 15.5.

El GSBPM también reconoce varios procesos globales<sup>2</sup> que se emplean a lo largo de las ocho fases. Éstos se pueden agrupar en dos categorías, los que tienen un componente estadístico y los procesos de soporte que son más generales y se pueden emplear en cualquier tipo de organización. Los del primer grupo se consideran más importantes en el contexto de este modelo, sin embargo también se deberán reconocer que los del segundo grupo (incluidos con detalle en el GAMSIO (UNECE 2019a)) tienen (a menudo de manera indirecta) impacto en varias partes del modelo.

Los procesos globales con una componente estadística incluyen los siguientes. Los cuatro primeros están más relacionados con el modelo.

- Gestión de calidad.- Este proceso incluye la evaluación de la calidad y mecanismos de control del proceso. Reconoce la importancia de la evaluación y el *feedback* a lo largo de todo el proceso estadístico;
- Gestión de metadatos.- Los metadatos son generados/reusados y procesados en cada fase, de donde se deriva una fuerte necesidad de disponer de un sistema de gestión de metadatos que asegure que los metadatos apropiados permanecen enlazados con los datos a lo largo de todo el GSBPM. Esto incluye consideraciones independientes del proceso tales como las figuras de custodio y responsable de los metadatos, su calidad, las reglas de archivo, preservación<sup>3</sup>, conservación<sup>4</sup> y eliminación;

---

<sup>2</sup>Traducimos *overarching* como global.

<sup>3</sup>Por preservación (*preservation*) se entiende el acto de conservar y mantener tanto la seguridad como la integridad de los datos. Se lleva a cabo mediante actividades formales gobernadas por políticas, regulaciones y estrategias dirigidas a proteger y prolongar la existencia y autenticidad de los datos y metadatos (Wikipedia 2021a).

<sup>4</sup>Por conservación (*retention*) definimos las políticas de gestión de datos y registros permanentes para cumplir los requisitos legales y de negocio de archivación de datos (Wikipedia 2021b).

- Gestión de datos - Esto incluye consideraciones independientes del proceso tales como la seguridad general de los datos, las figuras de custodio y responsable de los datos, la calidad de los datos, las reglas de almacenamiento, preservación, conservación y eliminación;
- Gestión de los datos de procesos.– Esto incluye las actividades de registro, sistematización y uso de los datos de la implementación del proceso de negocio.
- Gestión del conocimiento.– Asegura que los procesos estadísticos son repetibles, principalmente gracias al mantenimiento de la documentación del proceso;
- Gestión del marco estadístico.– Incluye desarrollar estándares, por ejemplo metodologías, conceptos y clasificaciones se utilizan de forma generalizada en múltiples procesos;
- Gestión del programa estadístico.– Incluye la monitorización sistemática y revisión de necesidades incipientes de información así como fuentes de datos nuevas y que cambian entre todos los dominios estadísticos. Puede dar como resultado la definición de nuevos procesos estadísticos o el rediseño de los existentes;
- Gestión de proveedores.– Incluye la gestión de la carga de procesos transversales, y temas como la caracterización y la gestión de la información de contacto (y por tanto está particularmente enlazada con procesos estadísticos que mantienen registros);
- Gestión de los usuarios.– Incluye actividades generales de marketing, promoción de los conocimientos estadísticos, y encargarse del *feedback* de usuarios no específicos.

Procesos globales más generales incluyen:

- Gestión de los recursos humanos;
- Gestión económica;
- Gestión de proyectos;
- Gestión del marco legal;
- Gestión del marco organizativo;
- Planificación estratégica.

### 15.2.2 Aplicabilidad

El objeto del GSBPM es su aplicación a todas las actividades llevadas a cabo por los productores de estadísticas oficiales, tanto a nivel nacional como a nivel internacional, con las que se obtienen resultados estadísticos. Se ha diseñado para ser independiente de la fuente de datos, por lo que puede ser utilizado para la descripción y evaluación de la calidad de procesos basados en encuestas, censos, registros administrativos, y otras fuentes no estadísticas o combinadas<sup>5</sup>.

---

<sup>5</sup>Esta afirmación, que aparece en el GSBPMv5.1, es discutible. Existen proyectos internacionales en el Sistema Estadístico Europeo que proponen arquitecturas de producción estadística que generalizan la

Mientras que los procesos estadísticos típicos incluyen la recogida y el procesamiento de los datos para producir resultados estadísticos, el GSBPM también se puede utilizar en casos en los que los datos existentes se revisan o que series de datos son recalculadas, tanto como resultado de una mejora en las fuentes de datos, como por un cambio en la metodología estadística. En estos casos, los datos de entrada son las estadísticas previamente publicadas, que son entonces procesadas y analizadas para producir resultados revisados. En tales casos, es probable que varios subprocesos y posiblemente algunas fases (particularmente las iniciales) sean omitidas. De forma similar, el GSBPM también se puede utilizar en procesos tales como la síntesis de Cuentas Nacionales y los típicos procesos de organizaciones estadísticas internacionales.

Además de usarse para procesos de los que se obtienen estadísticas, el GSBPM también se puede utilizar en el desarrollo y mantenimiento de registros estadísticos, donde los inputs son similares a los que se usan para la producción estadística (aunque típicamente poniendo el foco en datos administrativos), y los resultados son típicamente marcos u otras extracciones de datos, que serán entonces usados como inputs para otros procesos.

El GSBPM debe verse como un instrumento lo suficientemente flexible como para ser utilizado en todos los supuestos anteriores.

### 15.2.3 El uso del GSBPM

El GSBPM es un modelo de referencia. Está previsto que el GSBPM pueda ser utilizado por organizaciones a distintos niveles. Una organización puede elegir implementar el GSBPM de forma directa o usarlo como la base para desarrollar una adaptación específica del modelo.

Por ejemplo, el INE ha desarrollado un tercer nivel del GSBPM adaptado a las necesidades del Sistema Estadístico Nacional en España ([INE 2015](#)). Se puede usar en algunos casos sólo como un modelo al que las organizaciones se refieren en su comunicación interna o con otras organizaciones para clarificar discusiones. Los distintos escenarios para el uso del GSBPM son todos válidos.

Cuando las organizaciones hayan desarrollado adaptaciones específicas del GSBPM, pueden hacer algunas especializaciones al modelo para ajustarse a su contexto (véase p.ej. [INE 2015](#)). La evidencia hasta ahora sugiere que estas especializaciones no son suficientemente genéricas como para ser incluidas en el GSBPM.

En algunos casos puede ser apropiado agrupar algunos de los elementos del modelo. Por ejemplo, las fases uno a tres se puede considerar que corresponden a una única fase.

---

colección de funciones de negocio del GSBPM para abarcar necesidades de la producción derivadas del uso de *Big Data* (véase p.ej. [ESSnet on Big Data 2021](#)).

En otros casos, de forma particular para implementaciones prácticas, puede surgir la necesidad de añadir uno o más niveles detallados a la estructura indicada a continuación para identificar de manera separada diferentes componentes de los subprocesos.

### 15.3 Relaciones con otros modelos y estándares

Desde que se publicó el GSBPM, se han desarrollado varios modelos bajo el auspicio del HLG-MOS<sup>6</sup> para ayudar en la modernización de las estadísticas oficiales. De forma colectiva, se llama los modelos “ModernStats”. Los siguientes párrafos muestran los modelos ModernStats que están más relacionados con el GSBPM. Esta relación se muestra en la Figura 15.2

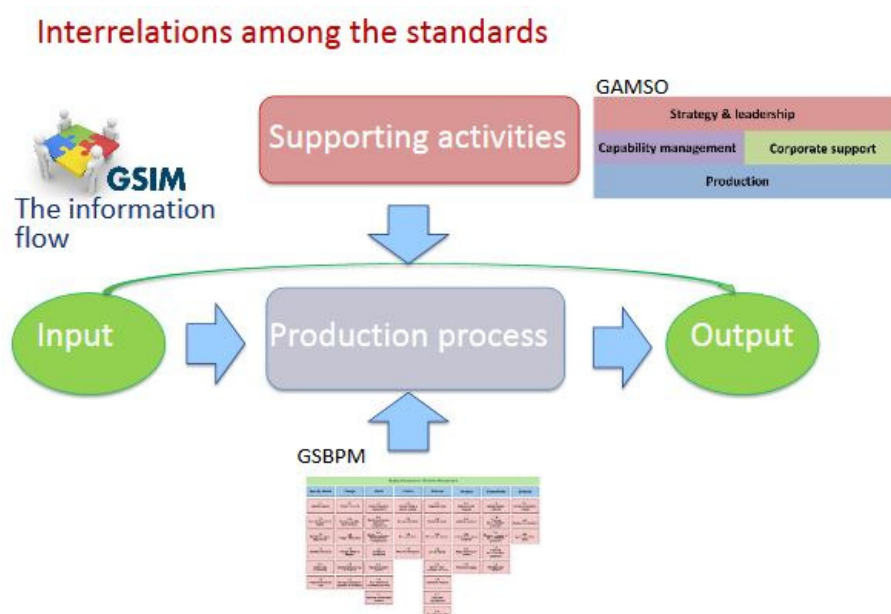


Figura 15.2: Relación entre GSIM, GSBPM y GAMSO.

#### 15.3.1 GAMSO

El GAMSO (UNECE 2019a) describe y define actividades que tienen lugar en una organización estadística. Extiende y complementa al GSBPM añadiendo actividades necesarias para la producción estadística (es decir actividades en las áreas de estrategia y dirección, desarrollo de capacidades y apoyo corporativo). En el GSBPM v5.0, algunas de estas actividades estaban incluidas como procesos generales. Las actividades que no están directamente relacionadas con la producción de estadísticas y/o son gestionadas a nivel corporativo o estratégico están ahora incluidas en el GAMSO (p.ej. gestión de los recursos humanos o las actividades de gestión de la calidad que se llevan a cabo a nivel corporativo como el desarrollo de un marco de calidad).

<sup>6</sup>High-Level Group for the Modernisation of Official Statistics.

El GAMS0 describe actividades – es decir, lo que las organizaciones estadísticas hacen. Incluye descripciones a alto nivel de estas actividades. Por un lado, el GSBPM se centra en el proceso de producción– describe más detalladamente cómo las organizaciones estadísticas llevan a cabo la actividad de producción estadística.

Como el GSBPM, el GAMS0 busca proporcionar un vocabulario común y un marco que fomente las actividades internacionales de colaboración. Se obtendrá un mayor valor del GAMS0 si se aplica de forma conjunta con el GSBPM.

### 15.3.2 GSIM

El GSIM (UNECE 2019b) es un marco de referencia para la **información estadística**, diseñado para jugar un papel importante en la modernización y racionalización de estadísticas oficiales tanto a nivel nacional como internacional. Facilita descripciones genéricas de la definición, gestión y uso de los datos y metadatos a lo largo del proceso de producción estadística. Proporciona un conjunto de objetos de información estandarizados y consistentemente descritos, que son los inputs y outputs en el diseño y producción de estadísticas.

El GSIM ayuda a explicar las relaciones significativas entre las entidades involucradas en la producción estadística y se puede usar para orientar el desarrollo y uso de estándares o especificaciones de implementación que resulten consistentes.

Como el GSBPM, el GSIM es uno de los pilares para la modernización de las estadísticas oficiales y alejarse del modelo actual de compartimentos estancos (*stovepipe*). El GSIM se ha diseñado para permitir enfoques innovadores en la producción estadística en la mayor medida posible; por ejemplo, en el área de la difusión, donde las demandas de agilidad e innovación están aumentando. También proporciona apoyo a aproximaciones vigente de la producción estadística.

El GSIM y el GSBPM son modelos complementarios para la producción y gestión de la información estadística. Como se muestra en la Figura 15.3, el GSIM ayuda a describir los subprocesos del GSBPM definiendo los objetos de información que fluyen entre ellos, que se crean en ellos, y que son usados por ellos para producir estadísticas oficiales. Los inputs y outputs se pueden definir en término de los objetos de información, y se formalizan en el GSIM.

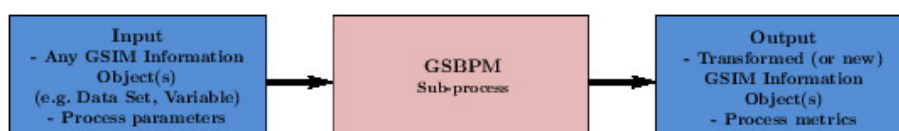


Figura 15.3: Relación entre GSIM y GSBPM en un paso de producción estándar.

Por tanto, se obtendrá un mayor valor del GSIM si se aplica de forma conjunta con el GSBPM. Del mismo modo, se obtendrá un mayor valor del GSBPM si se aplica de forma



conjunta con el GSIM. Sin embargo, es posible (aunque no ideal) aplicar el uno sin el otro.

De forma similar, ambos modelos apoyan la implementación de la *Common Statistical Production Architecture* (CSPA) (UNECE 2021), pero se puede aplicar con independencia de si se usa o no este marco de arquitectura de producción.

Aplicar de manera conjunta el GSIM y el GSBPM puede facilitar la construcción de sistemas eficientes gestionados con metadatos, y ayudar a armonizar infraestructuras informáticas estadísticas.

Una versión más desglosada de la Figura 15.3 se representa en la Figura 15.4, donde se observan tanto el flujo de datos como el flujo de metadatos<sup>7</sup>.

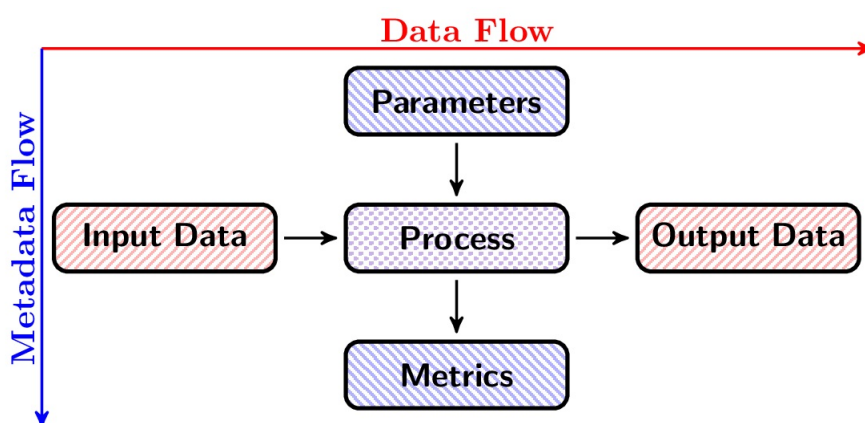


Figura 15.4: Relación entre GSIM y GSBPM en un paso de producción estándar desglosado.

## 15.4 Niveles 1 y 2 del GSBPM

Los niveles 1 y 2 del GSBPM están representados matricialmente en la Figura 15.5.

<sup>7</sup>Esta Figura está tomada de una conferencia de M. van der Loo en el INE en el año 2018 (véase Loo 2021, para sus referencias).

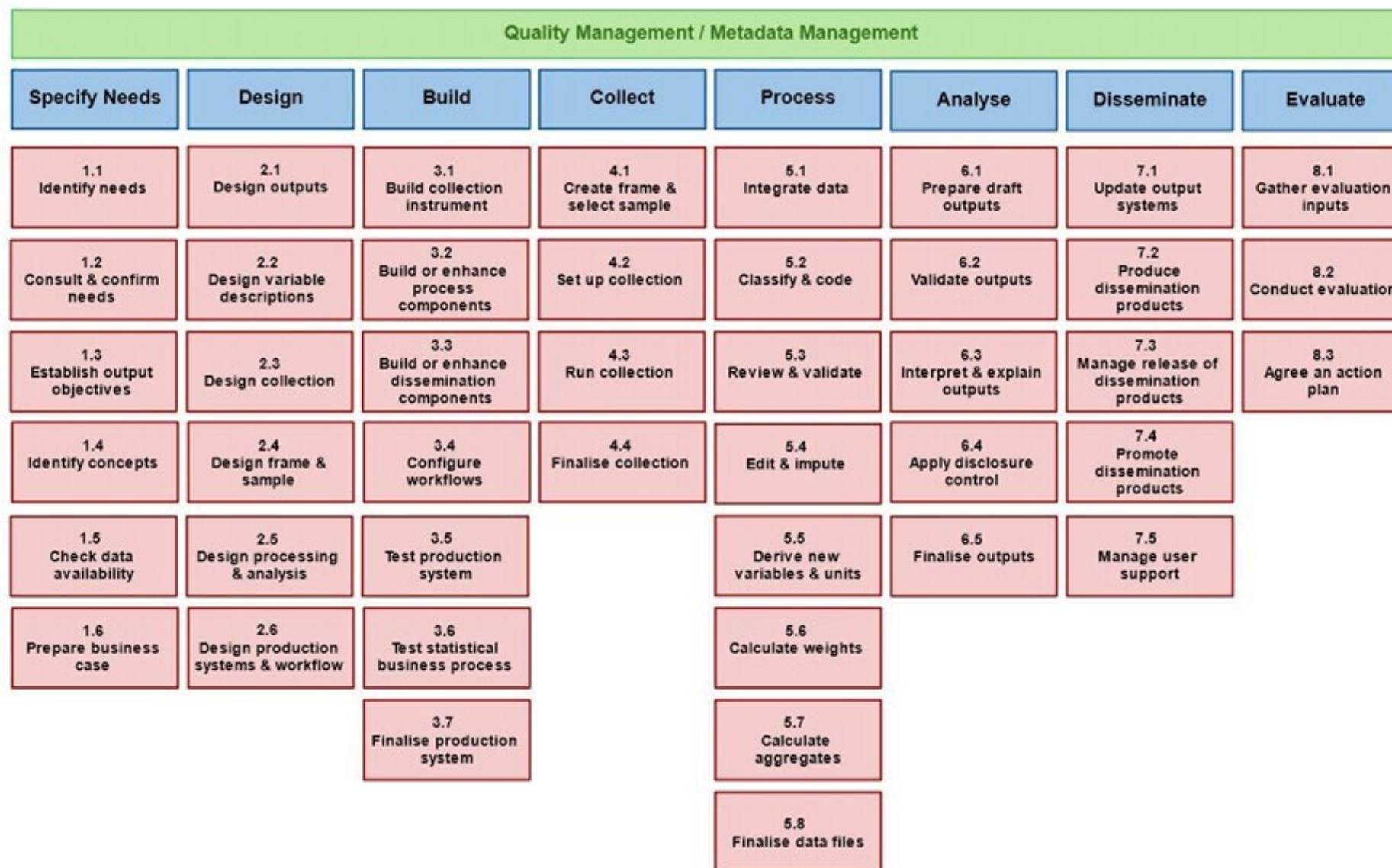


Figura 15.5: Niveles 1 y 2 del GSBPM



## 15.5 Descripciones de fases y subprocesos (fases 1 a 3)

A continuación se define cada fase, identificando los subprocesos dentro de cada fase, y describiendo sus contenidos.

### 1. Especificar necesidades

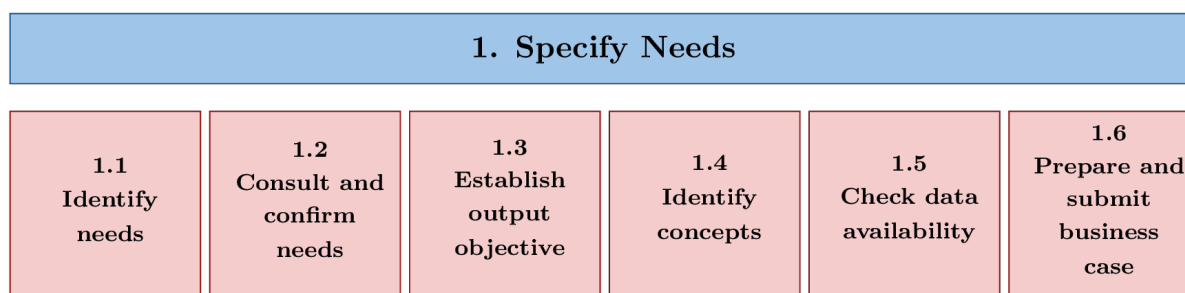


Figura 15.6: Fase 1 del GSBPM

Esta fase se desencadena cuando se detecta la necesidad de una nueva estadística o se inicia la revisión de una estadística existente como consecuencia de algún *feedback*. Incluye todas las actividades asociadas con la puesta en contacto con los usuarios para identificar de forma detallada sus necesidades estadísticas, proponiendo soluciones y preparando un proyecto que resuelva estas necesidades.

En esta fase la organización:

- identifica las necesidades de estadísticas;
- confirma, en mayor detalle, las necesidades de las distintas partes;
- establece los objetivos de los resultados estadísticos;
- identifica los conceptos y variables relevantes para los que se piden los datos;
- comprueba hasta qué punto las fuentes de datos actuales permiten alcanzar estas necesidades;
- prepara la documentación para la elaboración y justificación de la necesidad y viabilidad de un nuevo proyecto para conseguir la aprobación para producir las estadísticas.

Esta fase se divide en seis subprocesos. Estos son generalmente secuenciales, de izquierda a derecha, pero también pueden tener lugar en paralelo y pueden ser iterativos. Los subprocesos son:

#### 1.1. Identificar necesidades

Este subproceso incluye la investigación inicial e identificación de qué estadísticas son necesarias y qué se necesita de las estadísticas. Puede estar provocada por una nueva

solicitud de información o un cambio, como una reducción del presupuesto. Planes de acción provenientes de evaluaciones en iteraciones previas del proceso o de otros procesos relacionados pueden proporcionar un input a este subproceso. También incluye consideraciones sobre prácticas entre otras organizaciones estadísticas (nacionales e internacionales) que producen datos similares y, en particular, los métodos usados por esas organizaciones. Puede incluir consideraciones sobre las necesidades específicas de distintos tipos de usuarios, como los discapacitados, o distintos grupos étnicos.

### **1.2. Consultar y confirmar necesidades**

Este subproceso se centra en la consulta con las distintas partes interesadas y usuarios y confirma detalladamente las necesidades para producir las estadísticas en consideración. Se requiere un buen entendimiento de las necesidades de los usuarios para que la organización estadística sepa no solo qué es necesario publicar, sino también cuándo, cómo, y quizá, lo más importante, por qué. Para la segunda y posteriores iteraciones de esta fase, el principal objetivo será determinar si las necesidades previamente identificadas han cambiado. Este buen entendimiento de las necesidades de los usuarios es la parte crítica de este subproceso.

### **1.3. Establecer objetivos de los resultados**

Este subproceso identifica los resultados estadísticos que son necesarios para alcanzar las necesidades de los usuarios identificadas en el subproceso 1.2 (Consultar y confirmar necesidades). Incluye ponerse de acuerdo en la idoneidad de los resultados propuestos y sus medidas de calidad con los usuarios. Los marcos legales (p.ej. relacionados con la confidencialidad) y los recursos disponibles pueden ser limitaciones a la hora de establecer los objetivos.

### **1.4. Identificar conceptos**

Este subproceso clarifica los conceptos requeridos que serán medidos durante el proceso desde el punto de vista de los usuarios. En este punto los conceptos identificados pueden aún no estar alineados con estándares estadísticos existentes. Este alineamiento y la elección o definición de los conceptos y variables estadísticas que serán usadas tiene lugar en el subproceso 2.2.

### **1.5. Comprobar disponibilidad de datos**

Este subproceso comprueba si las fuentes actuales de datos podrían alcanzar los requisitos de los usuarios y las condiciones bajo las cuales estarían disponibles, incluyendo cualquier restricción en su uso. Una evaluación de posibles alternativas normalmente incluiría investigar potenciales fuentes de datos como los registros administrativos u fuentes otras no estadísticas para determinar si se podrían usar con fines estadísticos. Una vez que se han analizado las fuentes, se prepara una estrategia para cubrir las lagunas restantes. Este subproceso también incluye una evaluación más general sobre el marco legal en el cual se recogerán y usarán los datos y se pueda así identificar pro-

puestas de cambios en la legislación existente o la introducción de un nuevo marco legal.

### 1.6. Elaborar documentación para la elaboración y justificación de la necesidad y viabilidad de un nuevo proyecto

Este subproceso elabora documentación con los resultados de los otros subprocesos de esta fase en forma de caso de uso (*business case*) para evaluar y aprobar la implementación del proceso estadístico nuevo o modificado. Esta documentación necesitaría ajustarse a los requisitos de quien tiene que dar el visto bueno, pero generalmente incluye elementos como:

- Una descripción del proceso 'Como-Está'<sup>8</sup> proceso (si ya existe), con información sobre cómo se producen las estadísticas actuales, señalando las ineficiencias y aspectos a tener en cuenta;
- La solución propuesta para el proceso 'Futuro'<sup>9</sup>, detallando cómo el proceso estadístico se desarrollará para producir las estadísticas nuevas o revisadas;
- Una evaluación de los costes y beneficios así como cualquier restricción externa.

## 2. Diseñar

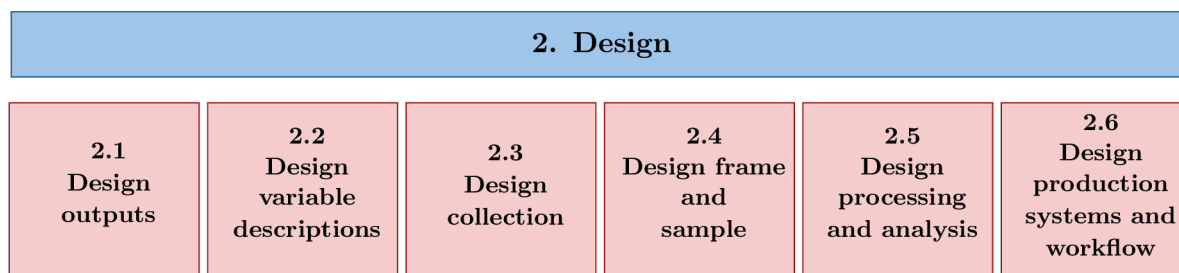


Figura 15.7: Fase 2 del GSBPM

Esta fase describe el desarrollo y actividades de diseño así como cualquier otro trabajo práctico de investigación asociados necesarios para definir los resultados estadísticos, conceptos, metodologías, instrumentos de recogida y procesos operacionales. Incluye todos los elementos del diseño necesarios para definir o refinar los productos estadísticos o servicios identificados en la documentación del caso de uso (subproceso 1.6.). Esta fase especifica todos los metadatos relevantes listos para su uso más tarde en el proceso estadístico, así como los procedimientos de garantía de calidad. Para resultados estadísticos producidos de forma periódica, esta fase normalmente tiene lugar la primera vez y cuando se identifiquen acciones de mejora en la fase de evaluación en alguna iteración previa.

Las actividades de diseño emplean estándares internacionales y nacionales con el fin de reducir la duración y el coste del diseño de proceso mejorando así la comparabilidad

<sup>8</sup>Traducimos *As-Is process* como proceso 'Como-Está'.

<sup>9</sup>Traducimos *To-Be process* como proceso 'Futuro'.

y usabilidad de los resultados. En este sentido, las organizaciones deben fomentar la reutilización o la adaptación de elementos de procesos existentes. Adicionalmente, los resultados de los procesos de diseño pueden formar la base de estándares futuros en la organización a nivel nacional o internacional.

Esta fase se divide en seis subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

### **2.1. Diseñar resultados**

Este subproceso contiene el diseño detallado de los resultados estadísticos, productos y servicios que se producirán, incluyendo los trabajos de desarrollo y preparación de los sistemas e instrumentos usados en la fase 'Difundir' (proceso 8.). Los métodos de control del secreto estadístico, así como los procesos relacionados con el acceso a los resultados confidenciales, también se diseñan en este subproceso. Los resultados deberían ser diseñados de forma que sigan los estándares existentes siempre que sea posible, por lo que los inputs de este proceso pueden incluir metadatos de operaciones de recogida de datos similares o anteriores, estándares internacionales e información sobre prácticas en otras organizaciones estadísticas relativas al subproceso 1.1 (Identificar necesidades).

### **2.2. Diseñar descripciones de variable**

Este subproceso define las variables estadísticas que se recogerán mediante los instrumentos de recogida, así como cualesquiera otras variables que se obtendrán a partir de ellas en el subproceso 5.5 (Derivar nuevas variables y unidades) y las clasificaciones estadísticas que se usarán. Se espera que se sigan, siempre que sea posible, los estándares nacionales e internacionales existentes. Puede ser necesario que este subproceso vaya en paralelo con el subproceso 2.3 (Diseñar recogida/obtención), mientras que la definición de las variables a recoger y la elección de los instrumentos de recogida pueden ser interdependientes hasta cierto punto. La preparación de la descripción de los metadatos de las variables recogidas y derivadas y las clasificaciones es una precondition necesaria para las fases posteriores.

### **2.3. Diseñar recogida**

Este subproceso determina los métodos e instrumentos de recogida más adecuados. Las actividades de este subproceso variarán de acuerdo con el tipo de instrumentos de recogida necesarios, que pueden incluir entrevistas asistidas por ordenador, cuestionarios en papel, interfaces con datos administrativos y técnicas de integración de datos. Este subproceso incluye el diseño de los instrumentos de recogida, preguntas y plantillas de respuesta (junto con las variables y clasificaciones estadísticas diseñados en el subproceso 2.2 (Diseñar descripciones de variable)). También incluye el diseño de cualquier acuerdo formal relacionado con el suministro de datos, tal como los memorandos de acuerdo y la confirmación de la base legal para la recogida de datos. Este subproceso se ve favorecido por herramientas tales como librerías de preguntas

(para facilitar la reutilización de preguntas y atributos relacionados), herramientas de diseño de cuestionarios (para posibilitar una compilación rápida y fácil de preguntas en formatos adecuados para tests cognitivos) y modelos de acuerdos (para ayudar a estandarizar términos y condiciones). Este subproceso también incluye el diseño de sistemas de gestión del informante específicos del proceso.

## **2.4. Diseñar marco y muestra**

Este subproceso solo se utiliza en procesos que involucren la recogida de datos basada en el muestreo. Este subproceso identifica y especifica la población de interés, define el marco muestral (y, donde sea necesario, el registro del que se deriva), y determina el criterio más apropiado de muestreo y la metodología (que puede incluir una enumeración completa). Las fuentes comunes para un marco muestral son los registros administrativos y estadísticos, los censos y la información de otras encuestas muestrales. Este subproceso describe cómo estas fuentes se pueden combinar si es necesario. Debe abordarse un análisis sobre si el marco cubre la población objetivo. Debe elaborarse un plan de muestreo. La muestra concreta se crea en el subproceso 4.1 (Crear marco y seleccionar muestra), usando la metodología especificada en este subproceso.

## **2.5. Diseñar procesamiento y análisis**

Este subproceso diseña la metodología del proceso estadístico que se aplicará durante las fases 'Procesar' y 'Analizar'. Puede incluir especificaciones de rutinas para codificar, depurar, imputar, estimar, integrar, validar y finalizar conjuntos de datos.

## **2.6. Diseñar sistemas de producción y flujos de trabajo**

Este subproceso determina los flujos de trabajo desde la recogida de datos hasta la difusión, con una visión total de todos los procesos necesarios en el conjunto del proceso de producción estadístico, y asegurando que todos juntos encajan de forma eficiente sin huecos o redundancias. A través del proceso se necesitan varios sistemas y bases de datos. Un principio general es la reutilización de procesos y tecnología entre muchos procesos estadísticos, por lo que las soluciones existentes en la producción (por ejemplo, servicios, sistemas y bases de datos) deberían de ser examinadas en primer lugar, para determinar si son adecuadas para este proceso específico, y, si se identifica alguna laguna, diseñar nuevas soluciones. Este subproceso también considera cómo el personal interactuará con los sistemas, y quién será responsable de qué y cuándo.

## **3. Desarrollar**

Esta fase construye y prueba las soluciones de producción hasta que estén listas para su uso en el entorno real de producción. Los resultados de la fase 'Diseñar' apunta a la selección de procesos reutilizables, instrumentos, información y servicios que estén ensamblados y configurados en esta fase para crear el entorno operacional completo para llevar a cabo el proceso. Nuevos servicios se desarrollan como una excepción, creados como respuesta a los vacíos en el catálogo actual de servicios existentes tanto

dentro de la organización como externamente. Estos nuevos servicios se construyen para que puedan ser ampliamente reutilizados dentro de la arquitectura de producción estadística.

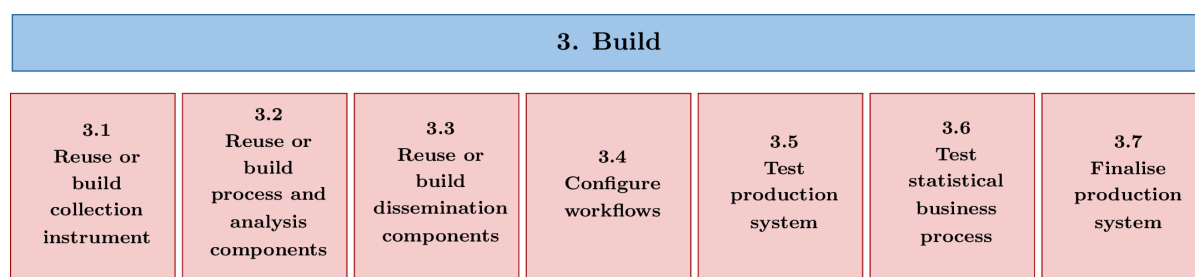


Figura 15.8: Fase 3 del GSBPM

Para los resultados estadísticos producidos de forma periódica, esta fase, más que en cada iteración, normalmente ocurre en la primera iteración y después de una revisión o un cambio metodológico o tecnológico.

Se descompone en siete subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo y pueden ser iterativos. Estos subprocesos son:

### 3.1. Desarrollar herramientas de recogida<sup>10</sup>

Este subproceso describe las actividades para construir los instrumentos de recogida que se usarán en la fase 'Recoger/Obtener'. El instrumento de recogida se genera o construye basado en las especificaciones de diseño creadas durante la fase 'Diseñar'. La recogida puede usar una o más formas de recogida, por ejemplo entrevistas personales o por teléfono; cuestionarios en papel, electrónicos o por web; centros de SDMX (*SDMX hubs*). Los instrumentos de recogida también pueden ser rutinas de extracción de datos usadas para combinar datos de conjuntos de datos estadísticos o administrativos ya existentes. Este subproceso también incluye preparar y probar los contenidos y funcionalidades del instrumento (por ejemplo, probar las preguntas en un cuestionario). Se recomienda considerar la conexión de los instrumentos de recogida con los sistemas de metadatos estadísticos, de forma que los metadatos se puedan recoger más fácilmente durante la fase de recogida. La conexión de los metadatos y los datos en el momento de recogida puede ahorrar trabajo en fases posteriores. La recogida de las métricas e indicadores de la recogida de datos (paradatos) también se considera importante en este subproceso.

### 3.2. Desarrollar o mejorar componentes de procesamiento

<sup>10</sup>Se toman los nombres de los subprocesos del GSBPM del estándar del INE (INE 2015).

Este subproceso describe las actividades para construir nuevos componentes y servicios o mejorar los existentes que son necesarios para las fases 'Procesar' y 'Analizar', tal y como están diseñados en la fase 'Diseñar'. Los servicios pueden incluir funciones y características para consolas de control (*dashboard*), servicios de información, funciones de transformación, entornos de flujos de trabajo y servicios de suministro y gestión de metadatos.

### 3.3. Desarrollar o mejorar componentes de difusión

Este subproceso describe las actividades para construir nuevos componentes y servicios o mejorar los existentes que son necesarios para la difusión de los productos estadísticos diseñados en el subproceso 2.1 (Diseñar resultados). Todos los tipos de componentes y servicios de difusión deben estar incluidos, desde los que se usan para producir publicaciones tradicionales en papel a los proporcionados por los servicios web, datos abiertos (*open data*) o accesos a microdatos.

### 3.4. Configurar flujos de trabajo

Este subproceso configura los flujos de trabajo, sistemas y transformaciones usadas en los procesos estadísticos, desde la recogida de datos a la difusión. Asegura que el flujo de trabajo especificado en el subproceso 2.6 (Diseñar sistemas de producción y flujo de trabajo) funciona en la práctica.

### 3.5. Probar sistemas de producción

Este subproceso está relacionado con la prueba de servicios montados y configurados así como con los flujos de trabajo relacionados. Incluye pruebas técnicas y la aprobación conjunta de los nuevos programas y rutinas, así como la confirmación de que las rutinas existentes en otros procesos estadísticos son adecuadas para su uso en este caso. Mientras que parte de esta actividad relacionada con la prueba de componentes y servicios individuales podría estar lógicamente enlazado con el subproceso 3.2 (Desarrollar o mejorar componentes de procesamiento), este subproceso también incluye la prueba de las interacciones entre servicios ensamblados y configurados y asegurar que las soluciones de producción funcionan como un conjunto coherente de procesos, información y servicios.

### 3.6. Probar procesos estadísticos

Este subproceso describe las actividades para gestionar pruebas de campo o pruebas piloto del proceso estadístico. Normalmente incluye una recogida a pequeña escala, para probar los instrumentos de recogida, seguido por el procesamiento y el análisis de los datos recogidos, para asegurar que el proceso estadístico funciona como se espera. Después del piloto, puede ser necesario volver a algún paso anterior y hacer ajustes en los instrumentos, sistemas o componentes. Para un proceso estadístico enorme, por ejemplo un censo de población, pueden ser necesarias varias iteraciones hasta que el proceso funciona de manera satisfactoria.

### 3.7. Finalizar sistema de producción

Este subproceso incluye las actividades para poner en marcha procesos y servicios montados y configurados, incluyendo servicios modificados y de nueva creación en producción listos para su uso. Las actividades incluyen:

- producir documentación sobre las componentes del proceso, incluyendo documentación técnica y manuales de usuarios;
- formación de los usuarios sobre cómo funciona el proceso;
- trasladar las componentes del proceso a un entorno de producción y asegurar que funcionan como se espera en tal entorno (esta actividad también puede ser parte del subproceso 3.5 (Probar sistema de producción)).

## Bibliografía

- ESSnet on Big Data (2021). *Work Package F on Process and Architecture*. Página visitada el día 28 de octubre de 2021. URL: [https://ec.europa.eu/eurostat/cros/content/WPF\\_Process\\_and\\_architecture\\_en](https://ec.europa.eu/eurostat/cros/content/WPF_Process_and_architecture_en).
- INE (2015). *Estándar de documentación de procesos de producción de operaciones estadísticas del INE: Los informes estandarizados de los metadatos de proceso*. Página visitada el día 28 de octubre de 2021. URL: [https://www.ine.es/clasifi/estandar\\_procesos.pdf](https://www.ine.es/clasifi/estandar_procesos.pdf).
- Loo, M. van der (2021). *Home Page*. URL: <http://www.markvanderloo.eu/>.
- UNECE (2019a). *Generic Activity Model for Statistical Organizations*. URL: <https://statswiki.unece.org/display/GAMSO/>.
- (2019b). *Generic Statistical Information Model v1.2*. URL: <https://statswiki.unece.org/display/gsim/>.
- (2019c). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.
- (2021). *Common Statistical Production Architecture*. URL: <https://statswiki.unece.org/display/CSPA>.
- Wikipedia (2021a). *Data Preservation*. Página visitada el día 28 de octubre de 2021. URL: [https://en.wikipedia.org/wiki/Data\\_preservation](https://en.wikipedia.org/wiki/Data_preservation).
- (2021b). *Data Retention*. Página visitada el día 28 de octubre de 2021. URL: [https://en.wikipedia.org/wiki/Data\\_retention](https://en.wikipedia.org/wiki/Data_retention).



## Tema 16

### Metadatos de la producción estadística. II. GSBPM. Descripciones de fases y subprocesos (fases 4 a 8). Procesos generales. Otros usos del GSBPM. Data Documentation Initiative (DDI), SDMX y comparación con el GSBPM.

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

UNECE (2019c). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>

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.

## 16.1 Descripciones de fases y subprocesos (fases 4 a 8)

### 4. Recoger/Obtener<sup>1</sup>

Esta fase recoge o recolecta toda la información necesaria (datos y metadatos), usando diferentes métodos de recogida (incluyendo la extracción de registros estadísticos, administrativos y otros no estadísticos así como bases de datos) y los carga en un entorno adecuado para un procesamiento posterior. Mientras que esto puede incluir la validación de los formatos de los conjuntos de datos, no incluye ninguna transformación de los datos en sí mismos, ya que todo esto se hace en la fase 'Procesar' (fase 5.). Para resultados estadísticos producidos de forma periódica, esta fase ocurre en cada iteración.

---

<sup>1</sup>Esta fase se denomina Recoger/Obtener. En lo que sigue se utilizará únicamente la palabra recoger. Obtener se refiere a las operaciones estadísticas en las que los datos no se recogen directamente de las unidades de la muestra mediante cuestionarios, sino que se obtienen a partir de registros administrativos o de otras fuentes de datos.

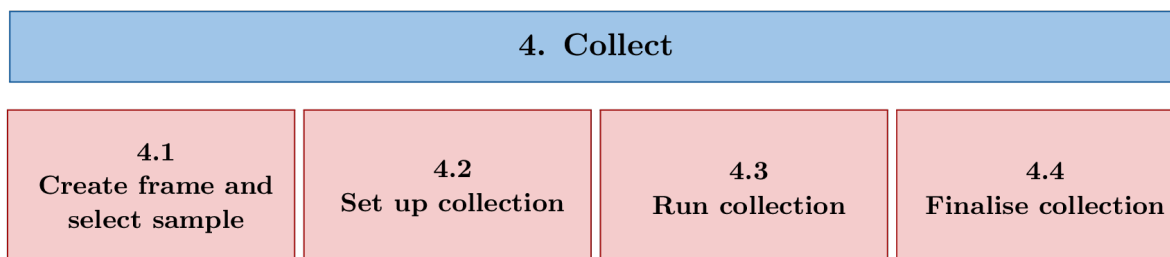


Figura 16.1: Fase 4 del GSBPM

La fase 'Recoger' se divide en cuatro subprocesos, que son generalmente secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo y pueden ser iterativos. Estos subprocesos son:

#### 4.1. Crear marco y seleccionar muestra

Este subproceso establece el marco y selecciona la muestra para esta iteración de la recogida, tal y como se especifica en el subproceso 2.4 (Diseñar marco y muestra). También incluye la coordinación de muestras entre repeticiones del mismo proceso estadístico (por ejemplo, para gestionar duplicidades o rotaciones) y entre diferentes procesos usando un marco o registro común (por ejemplo, para gestionar duplicidades o para distribuir la carga de respuesta). El control de calidad y la aprobación del marco y de la muestra seleccionada también tienen lugar en este subproceso, aunque el mantenimiento de registros subyacentes, de los cuales se extraen los marcos para varios procesos estadísticos, se trata como un proceso separado. Todo lo relacionado con la muestra en este subproceso no es normalmente relevante para procesos basados enteramente en el uso de fuentes preexistentes (por ejemplo, fuentes administrativas) ya que tales procesos generalmente crean marcos a partir de datos disponibles y después siguen una aproximación censal.

#### 4.2. Inicializar recogida/obtención

Este subproceso asegura que el personal, los procesos y la tecnología están listos para recoger los datos y metadatos a través de todos los métodos según la fase de diseño. Tiene lugar a lo largo de un período de tiempo que incluye la estrategia, planificación y las actividades de formación en preparación para la ejecución específica del proceso estadístico en el caso de uso en curso. En operaciones estadísticas periódicas, algunas (o todas) de estas actividades puede que no sean necesarias de forma explícita para cada iteración. Para procesos nuevos y únicos, estas actividades pueden llevar mucho tiempo. Este subproceso incluye:

- preparar una estrategia de recogida;
- formar al personal de recogida;
- asegurar que los recursos de recogida están disponibles (por ejemplo, ordenadores

portátiles o tablets);

- acordar los términos con cualquier organismo intermediario en la recogida (por ejemplo, subcontratos para las entrevistas por CATI – *Computer Assisted Telephone Interviewing*);
- configurar los sistemas de recogida para solicitar y recibir los datos;
- asegurar la seguridad de los datos recogidos;
- preparar los instrumentos de recogida (por ejemplo, imprimir cuestionarios, pre-rellenarlos con los datos existentes, cargar los cuestionarios y datos en los ordenadores de los entrevistadores, etc.).

Para fuentes de datos no provenientes de encuestas, este subproceso incluirá asegurar que existen los procesos, sistemas y procedimientos de confidencialidad necesarios dentro del proceso para recibir o extraer la información necesaria de la fuente.

#### 4.3. Ejecutar recogida/obtención

En este subproceso se implementa la recogida, junto con los diferentes instrumentos que se van a usar para recoger o recolectar la información, que puede incluir microdatos brutos o agregados producidos desde su origen, así como cualquier metadato asociado. Incluye el contacto inicial con informantes y cualquier seguimiento posterior o recordatorios. Puede incluir la entrada manual de datos durante el contacto con el informante o la gestión del trabajo de campo, dependiendo de la fuente y método de recogida. En este subproceso se graban cuándo y cómo se contacta con los informantes y si han contestado. Este subproceso también incluye la gestión de los informantes que forman parte de la recogida actual, asegurando que la relación entre la organización estadística y los informantes es positiva, grabando y respondiendo a los comentarios, peticiones y quejas. Para fuentes administrativas y otras fuentes no estadísticas, este proceso es breve: debe establecerse el contacto con el informante para que envíe la información, o la envía de acuerdo con un calendario fijado de antemano. Cuando la recogida alcanza su objetivo, se cierra y se produce un informe sobre el proceso. Algunas validaciones básicas de la estructura e integridad de la información recibida pueden tener lugar en este subproceso (por ejemplo, comprobar que los ficheros tienen el formato adecuado y contienen los campos esperados). Todas las validaciones sobre el contenido tienen lugar en la fase 'Procesar'.

#### 4.4. Finalizar recogida/obtención

Este subproceso incluye cargar los datos y metadatos recogidos en un entorno electrónico adecuado para un procesamiento posterior. Puede incluir la recogida de datos manual o automática, por ejemplo, usando personal administrativo o herramientas de reconocimiento óptico de caracteres para extraer información de los cuestionarios en papel o convertir los formatos de ficheros recibidos de otras organizaciones. También puede incluir el análisis de los metadatos y parados asociados con la recogida para asegurar que tales actividades han cumplido los requisitos exigidos. En operaciones estadísticas con un instrumento físico de recogida (como un cuestionario en papel) que

no se necesita para un procesamiento posterior, en este subproceso debe gestionarse el archivo de este material.

5. Procesar

Esta fase describe la depuración de los datos y su preparación para el análisis. Está formado por subprocesos que comprueban, depuran y transforman los datos de entrada de forma que puedan ser analizados y difundidos como resultados estadísticos. Se puede repetir varias veces si es necesario. Para resultados estadísticos producidos de manera periódica, esta fase tiene lugar en cada iteración. Los subprocesos en esta fase se pueden aplicar a datos tanto de fuentes estadísticas como no estadísticas (con la posible excepción del subproceso 5.6. Cálculo de pesos, que normalmente es específica de datos muestrales).

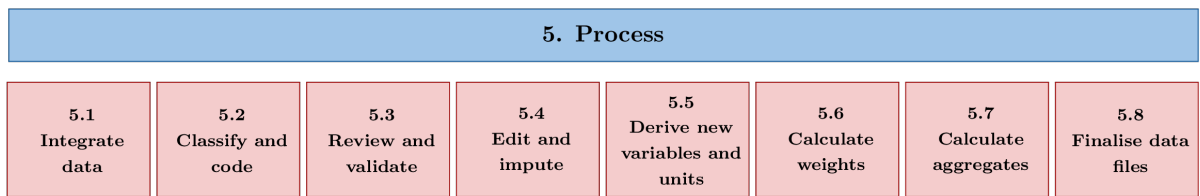


Figura 16.2: Fase 5 del GSBPM

Las fases ‘Procesar’ y ‘Analizar’ pueden ser iterativas y paralelas. El análisis (proceso 6.) puede revelar un conocimiento más amplio de los datos, que puede hacer evidente que sea necesario un procesamiento adicional en este proceso 5. Algunas actividades de las fases ‘Procesar’ y ‘Analizar’ pueden empezar antes de que la fase ‘Recoger/Obtener’ esté completada. Esto permite la recopilación de resultados provisionales cuando la oportunidad<sup>2</sup> es un requisito importante para los usuarios, incrementando así el tiempo disponible para análisis.

Esta fase se divide en ocho subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

5.1. Integrar datos

Este subproceso integra datos de una o más fuentes combinando los resultados de los subprocesos de la fase ‘Recoger/Obtener’. Los datos de entrada pueden venir de una mezcla de fuentes de datos externos e internos y de una variedad de métodos de recogida, incluyendo explotaciones de datos administrativos. El resultado es un conjunto de datos enlazados (*linked data*). La integración de datos puede incluir:

<sup>2</sup>Traducimos *timeliness* por oportunidad.

- combinar datos de múltiples fuentes como parte de la creación de estadísticas integradas (como las cuentas nacionales);
- rutinas de *matching* / *record linkage*, con el fin de unir micro o macro datos de distintas fuentes;
- priorizar, cuando dos o más fuentes contienen datos para la misma variable, con valores potencialmente diferentes.

La integración de datos puede tener lugar en cualquier momento de esta fase, antes o después de cualquier otro subproceso. También puede haber varias acciones de integración de datos a lo largo del proceso estadístico. Después de la integración, dependiendo de los requisitos de la protección de datos, los datos pueden ser anonimizados, es decir, eliminando los identificadores tales como el nombre y la dirección, para ayudar a proteger la confidencialidad.

## 5.2. Clasificar y codificar

Este subproceso clasifica y codifica los datos de entrada. Por ejemplo, rutinas de codificación automáticas (o manuales) pueden asignar códigos numéricos a respuestas de texto de acuerdo con un esquema de clasificación predeterminado.

## 5.3.-4. Revisar y validar. Depurar e imputar<sup>3</sup>

Este subproceso examina los datos para tratar de identificar problemas potenciales, errores y discrepancias tales como valores atípicos (*outliers*), falta de respuesta parcial y codificación errónea. También se puede referir a este subproceso como validación de los datos de entrada. Debe realizarse iterativamente, validando los datos de acuerdo con unas reglas de depuración predefinidas, normalmente en un orden preestablecido. Este subproceso puede señalar/marcar datos para la inspección o depuración automática o manual. La revisión y validación se pueden aplicar a datos de cualquier tipo de fuente, antes y después de la integración. Aunque la validación es tratada en este estándar como parte de la fase 'Procesar', en la práctica algunos elementos de la validación pueden tener lugar durante la ejecución de las actividades de recogida, particularmente para casos como la recogida web o la recogida asistida por ordenador. Aunque en el estándar internacional la detección de errores reales o potenciales se enmarca en el proceso 5.3, cualquier corrección que modifique los datos tiene lugar en el subproceso 5.4. En la adaptación del INE, ambas actividades figuran juntas porque es práctica común su ejecución integrada (detección+corrección).

Cuando los datos se consideren incorrectos, faltantes (*missing*) o de poca confianza/sin consistencia, en este subproceso se pueden incluir nuevos valores. Los términos depuración e imputación abarcan una gran variedad de métodos incluyendo enfoques basados en reglas. Pasos específicos generalmente incluyen:

---

<sup>3</sup>Aunque en la versión original del GSBPM los subprocesos 5.3 y 5.4 está separados, en el INE se ha decidido englobarlos en la adaptación del estándar ya que en la práctica, los procesos de revisión, validación, depuración e imputación se hallan altamente integrados (se ejecutan casi simultáneamente).

- la determinación sobre si incluir o cambiar datos;
- la selección del método a usar;
- añadir/cambiar los nuevos valores de datos;
- escribir los nuevos valores de datos en el conjunto de datos y marcarlos como cambiados;
- la producción de metadatos sobre el proceso de depuración e imputación.

### 5.5. Derivar nuevas variables y unidades

Este subproceso deriva datos para variables y unidades que no se obtienen de manera explícita en la recogida, pero que son necesarias para obtener los resultados estadísticos finales. Deriva nuevas variables aplicando una fórmula matemática a una o más variables que están presentes en el conjunto de datos, o aplicando distintos supuestos de un modelo. Esta actividad puede necesitar realizarse de manera iterativa, ya que algunas variables pueden a su vez estar basadas en otras variables derivadas. Por tanto, es importante asegurar que esas variables han sido derivadas en el orden correcto. Las nuevas unidades se pueden derivar agregando o dividiendo datos de unidades de recogida o mediante otros métodos de estimación. Algunos ejemplos incluyen obtener hogares cuando la unidad de recogida son las personas o empresas cuando la unidad de recogida son unidades jurídicas.

### 5.6. Calcular pesos

Este subproceso crea pesos para los datos de cada unidad de acuerdo con la metodología creada en el subproceso 2.5 (Diseño del proceso y análisis). En caso de una encuesta muestral, los pesos se pueden usar para 'elevar' los resultados para hacerlos representativos de la población objetivo, o para ajustar la falta de respuesta en censos. En otras situaciones, las variables pueden necesitar pesos con fines de normalización.

En la adaptación del INE, este subproceso se concentra en el cómputo de las ponderaciones para la construcción de índices compuestos a partir de índices simples para todos los grados de desagregación posible considerados en el diseño de la operación. El cómputo de pesos de muestreo calibrados, por el contrario, se considera una actividad altamente integrada con la construcción de estimadores y cálculo de agregados, por ello aparece en el proceso 5.7. Calcular agregados.

### 5.7. Calcular agregados

Este subproceso crea datos agregados y totales poblacionales a partir de microdatos o de agregados de nivel inferior. Incluye sumar datos de registros que comparten determinadas características, calculando medidas de promedios (media, mediana) y dispersión y aplicando los pesos del subproceso 5.6 para calcular índices compuestos adecuados. En el caso de una operación muestral, los errores muestrales (que involucra la estimación de varianzas) también se calculan en este subproceso asociándolos a los agregados correspondientes.

En la adaptación del INE, para operaciones estadísticas muestrales este subproceso incluye el cálculo de pesos de muestreo calibrados.

### 5.8. Finalizar ficheros de datos

Este subproceso integra los resultados de los otros subprocesos de esta fase y los resultados en un fichero de datos (normalmente de macrodatos), que se utiliza como el input de la fase 'Analizar'. Algunas veces puede ser un fichero intermedio más que uno final, particularmente para procesos de negocios en los que hay fuertes restricciones de tiempo y una necesidad de producir estimadores tanto preliminares como finales.

## 6. Analizar

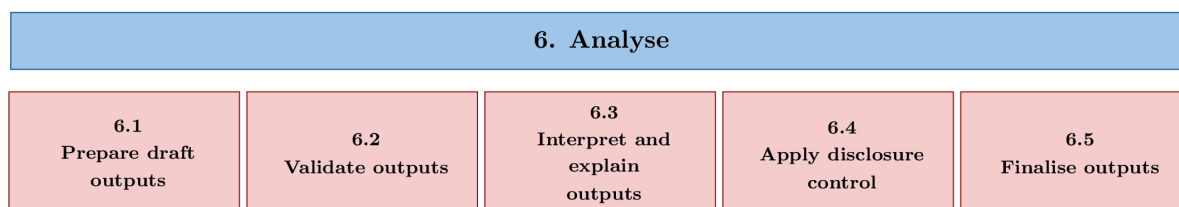


Figura 16.3: Fase 6 del GSBPM

En esta fase se producen resultados estadísticos que se examinan al detalle y se preparan para su difusión. Incluye preparar el contenido estadístico (incluyendo comentarios, notas técnicas, etc.), y asegurar que los resultados son 'adecuados para su propósito' antes de la difusión a los usuarios. Esta fase también incluye los subprocesos y actividades que permiten a los analistas estadísticos entender las estadísticas producidas. Para resultados estadísticos producidos periódicamente, esta fase ocurre en cada iteración. La fase 'Analizar' y sus subprocesos son genéricos para todos los resultados estadísticos, independientemente de la fuente que se haya utilizado.

La fase 'Analizar' se divide en cinco subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

### 6.1. Preparar borrador de resultados

En este subproceso se transforman los datos en resultados estadísticos. Incluye la producción de outputs como índices, tendencias o series ajustadas de estacionalidad, así como la documentación de los indicadores de calidad.

### 6.2.-3. Validar resultados. Interpretar y explicar los resultados<sup>4</sup>

<sup>4</sup>Aunque en la versión original del GSBPM los subprocesos 6.2 y 6.3 están separados, en el INE se ha decidido englobarlos ya que en la práctica, los procesos de validación, interpretación y explicación de resultados están altamente integrados.

En este subproceso los estadísticos evalúan y validan la calidad de los resultados producidos de acuerdo con el marco general de calidad y con las expectativas. Este subproceso también incluye actividades relacionadas con la recogida de información, con el efecto a largo plazo de construir un conjunto acumulado de conocimientos sobre un dominio estadístico específico. Este conocimiento se aplica a la recogida concreta, en el entorno concreto de producción en curso, para identificar cualquier divergencia de las expectativas y hacer posibles análisis bien fundamentados. Las actividades de validación pueden incluir:

- comprobar que la cobertura de la población y las tasas de respuesta son las requeridas;
- comparar los resultados con los ciclos previos (si es pertinente);
- comprobar que se dispone de los metadatos y parámetros (valores que adquieren los metadatos cuando se ejecuta el proceso de producción) asociados y que están en línea con las expectativas;
- confrontar los resultados con otros datos relevantes (tanto internos como externos);
- investigar las inconsistencias en los estadísticos, agregados, índices y tasas de variación;
- llevar a cabo la depuración macro;
- validar los estadísticos, agregados, índices y tasas con las expectativas y el dominio de conocimiento.

En este subproceso tiene lugar el entendimiento y comprensión en profundidad de los resultados por parte de los estadísticos. Los expertos usan este conocimiento para interpretar y explicar las estadísticas producidas para este ciclo evaluando en qué medida las estadísticas reflejan las expectativas iniciales, observando las estadísticas desde todas las perspectivas usando distintas herramientas y medios y llevando a cabo profundos análisis estadísticos.

#### **6.4. Aplicar control del secreto estadístico**

Este subproceso asegura que los datos (y metadatos) que se van a publicar no violan las reglas y normas legales de confidencialidad. Esto debe controlarse para comprobar el control del secreto estadístico primario y secundario, así como la aplicación de supresión de datos o las técnicas de perturbación. El grado y método de anonimización del control del secreto estadístico puede variar para distintos tipos de resultados, por ejemplo, el enfoque usado para conjuntos de microdatos con fines de investigación será diferente al que se usa para publicar tablas o mapas.

#### **6.5. Finalizar resultados**

Este subproceso asegura que las estadísticas y la información asociada son adecuados para el propósito y alcanzan el nivel de calidad requerido y, por tanto, están preparados



para su uso. Incluye:

- completar los controles de consistencia;
- determinar el nivel de difusión y aplicar excepciones;
- recopilar información complementaria, incluyendo interpretación, comentarios, notas técnicas, instrucciones/resúmenes, medidas de incertidumbre y cualquier otro metadato necesario;
- producir la documentación adicional interna;
- discutir previamente a la difusión con adecuados expertos internos en la materia;
- traducir los resultados a varios idiomas en países multilingües;
- aprobar el contenido estadístico para la difusión.

## 7. Difundir

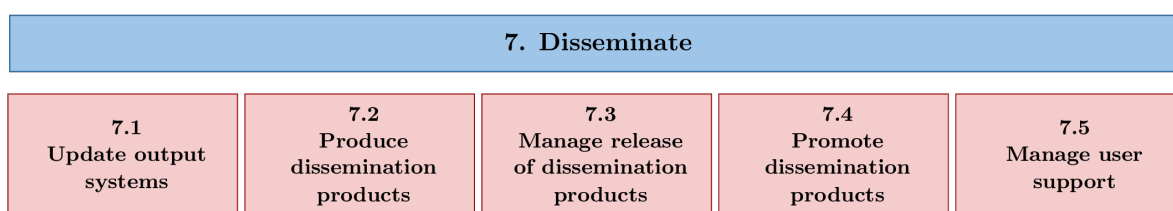


Figura 16.4: Fase 7 del GSBPM

Esta fase gestiona la difusión de los productos estadísticos para los usuarios. Incluye todas las actividades asociadas a la recopilación y publicación de un conjunto de productos estáticos y dinámicos vía una variedad de canales. Estas actividades ayudan al usuario en el acceso y uso de los resultados publicados por la organización estadística.

Para resultados estadísticos producidos de manera periódica, esta fase ocurre en cada iteración. Está formada por cinco subprocesos, que pueden ser secuenciales, de izquierda a derecha, pero también pueden ocurrir en paralelo, y pueden ser iterativos. Estos subprocesos son:

### 7.1. Actualizar sistemas de resultado

Este subproceso gestiona la actualización de los sistemas en los que los datos y metadatos se almacenan cuando están listos para su difusión, incluyendo:

- formatear datos y metadatos que están listos para poner en las bases de datos de resultados;
- cargar los datos y metadatos en las bases de datos de resultados;

- asegurar que los datos están enlazados con los metadatos correspondientes.

El formateo, la carga y el enlace de metadatos debe llevarse a cabo preferiblemente en fases anteriores, pero este subproceso incluye una revisión final de que todos los metadatos necesarios están en su sitio listos para la difusión.

## 7.2. Producir productos de difusión

Este subproceso produce los productos según el diseño previo (en el subproceso 2.1) para alcanzar las necesidades de los usuarios. Puede incluir publicaciones impresas, notas de prensa y páginas web. Los productos pueden tener muchas formas incluyendo gráficos interactivos, tablas, conjuntos de microdatos de uso público y ficheros que se pueden descargar. Los pasos típicos incluyen:

- preparar las componentes del producto (texto explicativo, tablas, gráficos, indicadores de calidad, etc.);
- juntar los componentes en productos;
- editar los productos y comprobar que alcanzan los estándares de publicación.

## 7.3. Gestionar divulgación de productos de difusión

Este subproceso asegura que existen todos los elementos para la difusión incluyendo la gestión del momento adecuado (*timing*) de la difusión. Incluye la difusión de la información para grupos específicos como la prensa o ministerios, así como acuerdos para embargos previos a la difusión. También incluye la provisión de productos a suscriptores y gestionar el acceso a datos confidenciales por parte de grupos de usuarios autorizados, como investigadores. Algunas veces una organización puede necesitar retractar un producto, por ejemplo, si se descubre un error. Esto también se incluye en este subproceso.

## 7.4. Promocionar productos de difusión

Mientras que el marketing en general se puede considerar un proceso global (*overarching process*), este subproceso está relacionado con una promoción activa de los productos estadísticos producidos en un proceso de negocio estadístico específico, de modo que se procure alcanzar la mayor audiencia posible. Incluye el uso de herramientas que gestionan la relación con los clientes, para dirigirse mejor a los potenciales usuarios de los productos, así como el uso de herramientas incluyendo páginas web, wikis y blogs para facilitar el proceso de comunicar la información estadística a los usuarios.

## 7.5. Gestionar soporte al usuario

Este subproceso asegura que las consultas y peticiones de servicios de los usuarios tales como el acceso a los microdatos se registran y procesan y que las respuestas se proporcionan dentro de las fechas límites acordadas. Estas consultas y peticiones deben ser revisadas de forma periódica para proporcionar un input al proceso global (*overarching process*) de control de calidad, ya que pueden indicar necesidades nuevas o cambiantes

de los usuarios.

## 8. Evaluar

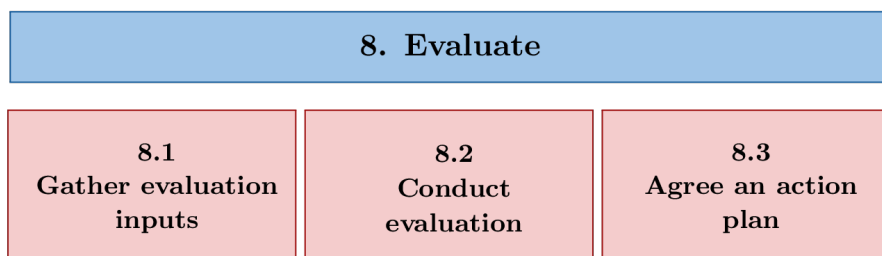


Figura 16.5: Fase 8 del GSBPM

Esta fase gestiona la evaluación de una ejecución específica del proceso, en contraposición con el proceso más general de gestión de la calidad estadística descrita en la sección siguiente. Lógicamente tiene lugar al final de la ejecución del proceso, pero depende de los inputs recogidos a lo largo de las diferentes fases. Incluye la evaluación del éxito de una ejecución específica del proceso de negocio estadístico, haciendo uso de un conjunto de inputs cuantitativos y cualitativos e identificando y priorizando mejoras potenciales.

Para resultados estadísticos producidos de manera periódica, la evaluación debe, por lo menos en teoría, tener lugar en cada iteración, determinando si futuras iteraciones deberían tener lugar, y en caso de que sea así, si debe implementarse alguna mejora. Sin embargo, en algunos casos, particularmente para procesos de negocio estadísticos periódicos bien consolidados, la evaluación puede no llevarse a cabo de manera formal en cada iteración. En tales casos, esta fase se puede ver como la forma de decidir si la próxima iteración debería empezar desde la fase 'Especificar necesidades' o desde algún fase posterior (a menudo la fase 'Recoger/Obtener').

Esta fase está compuesta por tres subprocesos, que generalmente son secuenciales, de izquierda a derecha, pero que se pueden solapar en la práctica. Estos subprocesos son:

### 8.1. Reunir inputs para la evaluación

El material para la evaluación puede producirse en cualquier otra fase o subproceso. Puede tomar muchas formas, incluyendo *feedback* de los usuarios, metadatos de proceso, métricas del sistema y sugerencias del personal. Los informes sobre el progreso de un plan de acción acordado en una iteración previa también pueden ser un input para evaluaciones de iteraciones posteriores. Este subproceso reúne todos estos inputs y los hace disponibles para la persona o equipo que lleve a cabo la evaluación.

### 8.2. Ejecutar evaluación

Este subproceso analiza los inputs de evaluación y los sintetiza en un informe de evaluación. El informe resultante debe tomar nota de cualquier asunto de calidad específico de esta iteración del proceso de negocio estadístico y debe hacer recomendaciones sobre cambios en caso de sea apropiado. Estas recomendaciones pueden abarcar cambios de cualquier fase o subproceso para futuras iteraciones del proceso o pueden sugerir que el proceso no se repita.

### 8.3. Acordar un plan de acción

Este subproceso ejecuta la toma de decisión necesaria para dar forma y acordar un plan de acción basado en el informe de evaluación. También debe de incluir consideraciones sobre un mecanismo para la supervisión del impacto de esas acciones, que pueden, a su vez, proporcionar un input para evaluaciones de futuras iteraciones del proceso.

## 16.2 Procesos generales (*overarching processes*)

El GSBPM también reconoce varios procesos generales que se aplican durante todas las fases de producción y a través de los procesos. Algunos de estos procesos generales se mencionan en el tema ???. Los procesos de gestión de la calidad y gestión de los datos y metadatos se explican más detenidamente en esta sección.

### Gestión de la Calidad

La calidad afecta a las organizaciones, procesos y productos. En el marco actual, el proceso general de gestión de la calidad se refiere a la calidad del producto y del proceso. La calidad a nivel institucional (por ejemplo, la adopción de una Política de Calidad o de un Marco de Garantía de Calidad) está incluido en el GAMSQ ([UNECE 2019a](#)).

El objetivo de la gestión de la calidad dentro del proceso estadístico es entender y gestionar la calidad de los productos estadísticos. Existe un acuerdo general entre las organizaciones estadísticas de que la calidad debería estar definida de acuerdo con la norma ISO 9000-2005 ([ISO9000:2005 2005](#)): 'Grado en el que un conjunto de características inherentes cumplen con los requisitos'. Por tanto, la calidad del producto es un concepto complejo y multifacético, normalmente definido en términos de varias dimensiones de calidad. Las dimensiones de la calidad que se consideran más importantes dependen de las perspectivas del usuario, de sus necesidades y prioridades, que varían entre procesos y grupos de usuarios.

Con el fin de mejorar la calidad del producto, la gestión de la calidad debe estar presente a lo largo de todo el modelo del proceso estadístico. Existe una estrecha relación con la Fase 8 (Evaluar), que tiene el papel específico de evaluar a posteriori casos y ejecuciones individuales de procesos estadísticos. Sin embargo, la gestión de la calidad tiene una cobertura más profunda y más amplia. Así como la evaluación de las iteraciones de un

proceso, también es necesario evaluar de forma separada las fases y subprocesos, lo óptimo sería cada vez que se utilizan, pero por lo menos de acuerdo con un calendario acordado. Los metadatos generados por los distintos subprocesos en sí mismos también resultan de interés como input para la gestión de la calidad del proceso. Estas evaluaciones se pueden implementar en procesos específicos o entre varios procesos que usen componentes comunes.

Además, el conjunto de acciones que deben implementarse en los subprocesos para evitar y controlar los errores juegan un papel fundamental en la gestión de la calidad. La estrategia se puede incluir en el plan de garantía de la calidad.

Dentro de una organización, la gestión de la calidad se referirá generalmente a un marco específico de calidad y, por tanto, puede tomar distintas formas y ofrecer distintos resultados en diferentes organizaciones. La multiplicidad de marcos de calidad existentes aumenta la importancia de los estudios comparativos y las evaluaciones mediante revisión por pares (*peer review*) y aunque resulta dudoso que estos enfoques sean factibles para cada iteración de cada parte de cada proceso estadístico, debe usarse de forma sistemática de acuerdo con un calendario acordado que permita la revisión de las principales partes del proceso en un período de tiempo específico.

Ampliando el campo de aplicación del proceso general de gestión de la calidad, también se puede considerar la evaluación de grupos de procesos estadísticos con el fin de identificar potenciales duplicaciones o lagunas.

Todas las evaluaciones darán lugar a *feedback*, que debe usarse para mejorar el proceso, fase o subproceso relevante, creando un bucle de calidad.

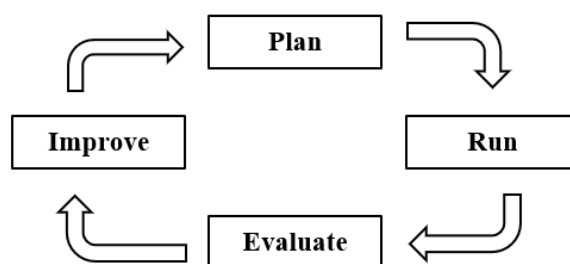


Figura 16.6: Bucle de calidad

Ejemplos de actividades de gestión de la calidad incluyen:

- Establecer y mantener el marco de calidad;
- Establecer criterios globales de calidad;
- Establecer los objetivos de calidad del proceso y supervisar su cumplimiento;
- Solicitar y analizar el *feedback* de los usuarios;

- Revisar las acciones y la documentación de las lecciones aprendidas;
- Examinar los metadatos de proceso y los indicadores de calidad;
- Auditorías internas o externas de procesos estadísticos.

Los indicadores de calidad sirven de ayuda a la gestión de calidad orientada a los procesos. Una lista de indicadores de calidad para las fases y subfases del GSBPM así como para los procesos generales de gestión de la calidad y de los metadatos se puede encontrar en los Indicadores de Calidad del GSBPM (UNECE 2018). Entre otros, se pueden usar para identificar carencias y/o duplicidades de trabajo en la organización.

### Gestión de los metadatos

Los metadatos juegan un papel muy importante y deben ser gestionados a un nivel operativo dentro del proceso de producción estadística. Cuando los aspectos de la gestión de los metadatos se consideran a nivel corporativo o estratégico (p.ej. hay sistemas de metadatos que afectan a grandes partes del sistema de producción) deberían ser considerados en el GAMS (UNECE 2019a).

Una buena gestión de los metadatos es esencial para un eficiente funcionamiento del proceso estadístico. Los metadatos están presentes en cada fase, tanto si son creados como si son transferidos de una fase previa. En el contexto de este modelo, el énfasis del proceso general de gestión de los metadatos está en la creación, uso y archivo de metadatos estadísticos, aunque los metadatos de los distintos subprocesos en sí mismos también son de interés, incluyéndolos como un input para la gestión de calidad. El principal problema es asegurar que estos metadatos son recogidos tan pronto como sea posible y almacenados y transferidos de una fase a otra junto con los datos a los que se refieren. La estrategia y los sistemas de gestión de los metadatos son, por tanto, vitales para el funcionamiento de este modelo y pueden ser facilitados por el GSIM (UNECE 2019b).

El GSIM es un marco de referencia de objetos de información que permite descripciones genéricas de la definición, gestión y uso de datos y metadatos durante todo el proceso de producción. El GSIM favorece un enfoque consistente para los metadatos, facilitando el papel básico de los metadatos, es decir, que estos deben definir de forma única y formalmente el contenido y los enlaces entre los objetos de información y los procesos de producción en el sistema de información estadística.

El Marco Común de Metadatos METIS (*METIS Common Metadata Framework*) (UNECE 2021b) identifica los siguientes dieciséis principios esenciales para la gestión de los metadatos, todos ellos previstos que se cubran en el proceso general de gestión de los metadatos y tenidos en consideración cuando se diseñe e implemente el sistema de metadatos estadísticos. Los principios se presentan en cuatro grupos:

Tratamiento de los metadatos	
	<ul style="list-style-type: none"> <li>i. Modelo del Proceso de Negocio Estadístico: Gestiónense los metadatos con atención al conjunto del modelo de proceso estadístico.</li> <li>ii. Activo no pasivo: Háganse los metadatos activos en la mayor medida posible. Metadatos activos son metadatos que conducen a otros procesos y acciones. Tratando los metadatos de esta forma se asegurará que son acurados y actualizados.</li> <li>iii. Reutilización: Reutilícense los metadatos donde sea posible para la integración estadística así como por razones de eficiencia.</li> <li>iv. Versiones: Presérvese la historia (versiones anteriores) de los metadatos.</li> </ul>
Autoridad de metadatos	
	<ul style="list-style-type: none"> <li>i. Registro: Asegúrese que el proceso de registro (flujo de trabajo) asociado con cada elemento de los metadatos está bien documentado de forma que esté clara la identificación del titular, estado de aprobación, fecha de la operación, etc.</li> <li>ii. Fuente única: Asegúrese que existe una única fuente, autorizada ('autoridad de registro') para cada elemento de los metadatos.</li> <li>iii. Una entrada/actualización: Minimícense los errores con una entrada de datos única y realizando actualizaciones en un único sitio.</li> <li>iv. Variaciones de los estándares: Asegúrese que las variaciones de los estándares son estrictamente gestionadas/aprobadas, documentadas y visibles.</li> </ul>
Relación con el Ciclo/Proceso Estadístico	
	<ul style="list-style-type: none"> <li>i. Integridad: Hágase del trabajo relacionado con los metadatos una parte integral de los procesos en toda la organización.</li> <li>ii. Enlace de metadatos: Asegúrese que los metadatos presentados a los usuarios enlazan con los metadatos que condujeron el proceso o que se crearon durante el proceso.</li> <li>iii. Describir el flujo: Describase el flujo de metadatos con los procesos de negocios y estadístico (junto con el flujo de datos y la lógica de negocio).</li> <li>iv. Captura en la fuente: Captúrense los metadatos en su fuente, preferiblemente de forma automática como un subproducto de otros procesos.</li> <li>v. Intercambio y uso: Intercámbiense los metadatos y úsense para proporcionar información tanto para procesos informatizados automáticos como para la interpretación humana. La infraestructura para intercambio de datos y sus metadatos asociados debería estar basada en componentes sin conexión directa, con la opción de lenguajes estándares de intercambio, como el XML.</li> </ul>
Usuarios	
	<ul style="list-style-type: none"> <li>i. Identifica usuarios: Asegúrese que los usuarios se identifican con claridad para todos los procesos relativos a los metadatos y que todos los metadatos capturados generarán valor para ellos.</li> <li>ii. Distintos formatos: La diversidad de metadatos debe reconocerse de modo que haya distintos puntos de vista correspondientes con los distintos usos de los datos. Los distintos usuarios requieren distintos niveles de detalle. Los metadatos aparecen en diferentes formatos dependiendo de los procesos y objetivos para los que se han producido y usado.</li> <li>iii. Disponibilidad: Asegúrese que los metadatos están disponibles pronto y aprovechables en el contexto de las necesidades de información de los usuarios (tanto si el usuario es interno o externo).</li> </ul>

## Gestión de los datos

La gestión de datos es esencial ya que los datos se producen en muchas de las activida-

des de los procesos y son los principales outputs. El principal objetivo de la gestión de los datos es asegurar que los datos se usan de forma apropiada y son útiles a lo largo de su ciclo de vida. La gestión de los datos a lo largo de su ciclo de vida incluye actividades como la planificación y la evaluación de los procesos de gestión de los datos así como la creación y la implementación de procesos relacionados con la recogida, organización, uso, protección, preservación y eliminación de los datos.

Cómo los datos son gestionados estará muy vinculado al uso de los datos, que por otro lado está relacionado con los procesos estadísticos donde los datos son creados. Tanto los datos como los procesos en los que son creados deben estar bien definidos con el fin de asegurar una gestión adecuada de los datos.

Ejemplos de actividades de gestión de los datos incluyen:

- Establecer una estructura de gobierno y asignar responsabilidades administrativas de los datos;
- Diseñar estructuras de datos y sus conjuntos de datos asociados y el flujo de datos a lo largo del proceso estadístico;
- Identificar bases de datos (repositorios) para almacenar los datos y la gestión de la base de datos;
- Documentar los datos (p.ej. registrando e inventariando los datos, clasificando los datos de acuerdo con su contenido, permanencia u otras clasificaciones necesarias);
- Determinar los periodos de conservación de los datos<sup>5</sup>;
- Asegurar los datos frente al acceso y uso no autorizado;
- Proteger los datos frente a cambios tecnológicos, deterioro de los medios físicos y corrupción de datos;
- Realizar comprobaciones de la integridad de los datos (p.ej. comprobaciones periódicas que aseguren la acuracidad y consistencia de los datos a lo largo de su ciclo de vida);
- Realizar actividades de tratamiento una vez que el periodo de conservación de los datos ha expirado.

## 16.3 Otros usos del GSBPM

El objetivo original del GSBPM era proporcionar una base para que las organizaciones estadísticas acordaran una terminología estándar para ayudar en sus discusiones sobre el desarrollo de un sistema de metadatos estadísticos y de procesos. Sin embargo, a medida que el modelo se ha desarrollado, ha resultado cada vez más aparente que se

---

<sup>5</sup>Por conservación (*retention*) definimos las políticas de gestión de datos y registros permanentes para cumplir los requisitos legales y de negocio de archivación de datos ([Wikipedia 2021](#)).



puede usar para muchos otros propósitos, en particular, relacionados con la modernización de estadísticas oficiales. Los artículos y documentos que describen los usos actuales y potenciales del GSBPM están disponibles en la wiki de la UNECE ([UNECE 2021a](#)). La lista que figura a continuación tiene por objetivo señalar algunos usos actuales e inspirar nuevas ideas sobre cómo el GSBPM se puede usar en la práctica.

- Proporcionar una estructura para la documentación de procesos estadísticos.- El GSBPM puede proporcionar una estructura para organizar y acumular la documentación dentro de una organización, promocionando la estandarización y la identificación de buenas prácticas;
- Facilitar la compartición de los métodos estadísticos y el software.- El GSBPM denie los componentes de los procesos estadísticos de modo que no solo fomenta el intercambio de métodos y herramientas informáticas, sino que también facilita el intercambio entre distintas organizaciones estadísticas que utilizan el modelo;
- Describir los estándares que son utilizados o podrían ser usados para las distintas fases del proceso estadístico. Por ejemplo, el Anexo 2 de la guía de usuario del estándar SDMX 2.1 ([SDMX 2012](#)) explora cómo el SDMX se usa en el trabajo estadístico en el contexto de un modelo de procesos de negocio;
- Proporcionar un marco para la evaluación y mejora de la calidad del proceso.- Si un enfoque por comparación a la evaluación de la calidad del proceso resulta exitoso, es necesario estandarizar procesos tanto como sea posible. El GSBPM proporciona un mecanismo para facilitar esto;
- Integrar mejor los trabajos sobre metadatos y calidad estadísticos.- Enlazado con el punto anterior, el marco común proporcionado por el GSBPM puede ayudar a integrar el trabajo internacional sobre los metadatos estadísticos con aquél sobre calidad de datos proporcionando un marco y una terminología comunes para describir el proceso estadístico;
- Proporcionar el modelo básico para marcos estándares metodológicos.- Los estándares metodológicos se pueden enlazar con la(s) fase(s) o subproceso(s) con los que están relacionados y pueden entonces ser clasificados y almacenados en una estructura basada en el GSBPM;
- Desarrollar un modelo de repositorio de proceso de negocio para almacenar outputs del modelado de procesos y enlazarlos con el modelo de proceso de negocio estadístico;
- Proporcionar un marco subyacente para desarrollar y un conjunto de terminología estándar para describir competencias y *expertise* necesarios en el proceso de producción estadística;
- Medir los costes operacionales.- El GSBPM se puede usar como una base para medir el coste de las distintas partes de un proceso estadístico. Esto ayuda a localizar actividades de modernización que mejoren la eficiencia de las partes del proceso que son más costosas;
- Medir el rendimiento del sistema.- Relacionado con el punto anterior sobre los

costes, el GSBPM también se puede usar para identificar componentes que no se están realizando de forma eficiente, que se están duplicando una a otra innecesariamente o que necesitan ser reemplazadas. De forma similar pueden identificarse carencias para las cuales se deben desarrollar nuevos componentes;

- Proporcionar una herramienta para alinear procesos de negocio de proveedores de datos no estadísticos (p.ej. datos administrativos o geoespaciales) facilitando la comunicación entre estadísticos y expertos de otros dominios y para armonizar la terminología relacionada;
- Proporcionar una herramienta para fortalecer la capacidad y el conocimiento técnico metódicamente mediante referencias detalladas a cada fase de producción;
- Proporcionar una herramientas para el desarrollo y revisión de las clasificaciones estadísticas.

## 16.4 Data Documentation Initiative (DDI), SDMX y comparación con el GSBPM

### El modelo combinado del ciclo de vida DDI 3

Este modelo se ha desarrollado en el consorcio *Data Documentation Initiative* (DDI) ([DDI Alliance 2021](#)), una iniciativa internacional para establecer un estándar de documentación técnica para describir los datos en ciencias sociales. La Alianza DDI incluye principalmente instituciones académicas y de investigación, por tanto, el alcance de este modelo es un poco distinto del GSBPM, que se aplica de manera específica a las organizaciones de estadística oficial. A pesar de esto, el proceso estadístico parece bastante similar entre los productores de estadísticas oficiales y no oficiales, como puede observarse en la consistencia a alto nivel entre los modelos.

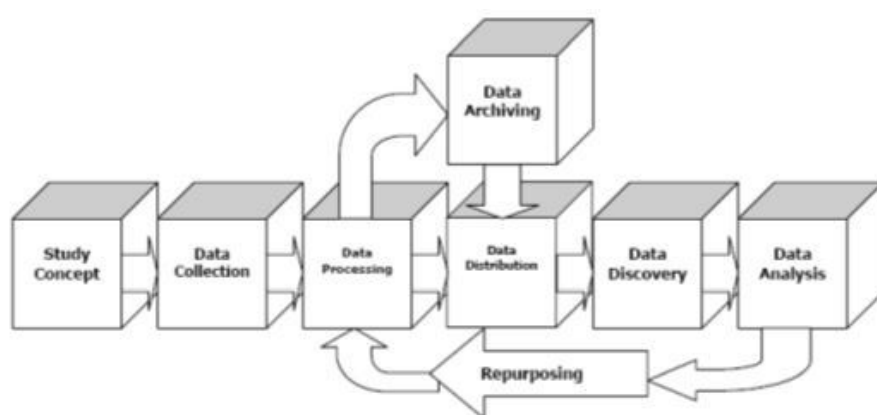


Figura 16.7: El modelo combinado del ciclo de vida DDI 3

Las principales diferencias entre el GSBPM y el modelo combinado del ciclo de vida DDI son:

- El GSBPM generalmente asume que el proceso general es llevado a cabo por una única organización (aunque para procesos grandes como los censos algunos

subprocesos pueden externalizarse). El modelo DDI parece reconocer que algunos pases como 'Análisis de datos' y 'Reutilización' pueden ser llevados a cabo por organizaciones distintas de la que recoge los datos. Esto refleja una diferencia fundamental entre las prácticas en las comunidades investigadora y de estadísticas oficiales, donde la comunidad investigadora tiene mayores posibilidades de colaboración entre organizaciones durante el proceso de producción.

- El modelo DDI reemplaza la fase de difusión por 'Distribución de Datos' que tiene lugar antes de la fase de análisis. Esto refleja la diferencia en enfoque entre las comunidades investigadora y de estadísticas oficiales, donde la última pone un mayor énfasis en la difusión de los datos, más que en una investigación basada en la difusión por parte de terceros.
- El modelo DDI contiene el proceso de 'Reutilización', definido como el uso secundario de un conjunto de datos o la creación de un conjunto de datos real o virtual armonizado. Esto generalmente se refiere a algún reuso de un conjunto de datos que no fue originalmente previsto en las fases de diseño y recogida. En el GSBPM, si los outputs de algún proceso son reutilizados para algún otro propósito, se tratan en dos procesos separados (dos casos del modelo). El segundo proceso identifica los datos en la fase 1 (Especificar Necesidades) donde hay un subproceso para confirmar la disponibilidad de datos existentes, y los obtiene en la fase 4 (Recoger/Obtener) y luego los usa para producir nuevos outputs.
- El modelo DDI tiene fases separadas para la localización de los datos y el análisis de datos, mientras que en el GSBPM estas funciones están combinadas en la fase 6 (Analizar). En algunos casos, los elementos de la fase de análisis del GSBPM también pueden estar incluidas en la fase 'Procesamiento de Datos' del DDI, dependiendo de la extensión del trabajo analítico anterior a la fase 'Distribución de Datos'.
- El GSBPM identifica de forma explícita 'procesos generales', como la calidad y la gestión de metadatos, mientras que éstos son más implícitos en el modelo DDI.

Está claro que el GSBPM y el modelo combinado del ciclo de vida DDI sirven para propósitos ligeramente diferentes. Esto proporciona una buena justificación para las diferencias entre ellos. Sin embargo, a pesar de estas diferencias en su finalidad también hay un montón de similitudes. Por tanto, es útil establecer una correspondencia entre los dos modelos para intentar tener un mejor conocimiento de cómo interactúan.

## SDMX

El estándar SDMX (*Statistical Data and Metadata eXchange*) ([SDMX 2012](#)) no proporciona un modelo para procesos estadísticos en el mismo sentido que el GSBPM. Pero proporciona una terminología estándar para datos y metadatos estadísticos, así como estándares técnicos y directrices orientadas hacia el contenido para la transferencia de datos y metadatos, que se podrían aplicar entre subprocesos dentro de una organización estadística. Por tanto, se considera que se puede incorporar el GSBPM en las directrices orientadas hacia el contenido del SDMX como un dominio transversal. También se

considera que el SDMX puede proporcionar el formato para la transmisión de datos entre subprocesos dentro de una organización estadística.

## Bibliografía

- DDI Alliance (2021). *Data Documentation Initiative*. URL: <https://ddialliance.org/>.
- ISO9000:2005 (2005). *Sistemas de gestión de la calidad*. URL: <https://www.iso.org/obp/ui/es/#iso:std:iso:9000:ed-4:v1:es>.
- SDMX (2012). *SDMX 2.1 User Guide*. URL: [https://sdmx.org/wp-content/uploads/SDMX\\_2-1\\_User\\_Guide\\_draft\\_0-1.pdf](https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf).
- UNECE (2018). *Quality Indicators for the GSBPM*. URL: <https://statswiki.unece.org/display/GSBPM/Quality+Indicators>.
- (2019a). *Generic Activity Model for Statistical Organizations*. URL: <https://statswiki.unece.org/display/GAMSO/>.
  - (2019b). *Generic Statistical Information Model v1.2*. URL: <https://statswiki.unece.org/display/gsim/>.
  - (2019c). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.
  - (2021a). *Statistics Wiki*. URL: <https://statswiki.unece.org/>.
  - (2021b). *The Common Metadata Framework*. URL: <https://statswiki.unece.org/display/hlgbas/The+Common+Metadata+Framework>.
- Wikipedia (2021). *Data Retention*. Página visitada el día 28 de octubre de 2021. URL: [https://en.wikipedia.org/wiki/Data\\_retention](https://en.wikipedia.org/wiki/Data_retention).

## Tema 17

**Metadatos de la producción estadística. III. GSIM. Introducción. Alcance. ¿Qué es el GSIM? Beneficios del GSIM para la organización como un todo. GSIM y GSBPM. ¿Qué implica para el estadístico?: Puntos de vista del negocio y de la tecnología de la información. SDMX, DDI y otros estándares.**

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

UNECE (2019b). *Generic Statistical Information Model v1.2*. URL: <https://statswiki.unece.org/display/gsim/>

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.

### 17.1 Introducción genérica al GSIM

El Modelo Genérico de Información Estadística (*Generic Statistical Information Model - GSIM*) (UNECE 2019b) es un estándar de producción estadística centrado en la modelización de la información contenida en el proceso de producción en una oficina de estadística. Está constituido por un corpus de documentación a distintos niveles y, por tanto, dirigido a distinta audiencia dentro de la producción estadística oficial. La UNECE recomienda los documentos por especialización técnica y/o responsabilidades según la Tabla 17.1.

De toda esta documentación, este tema básicamente reproduce el *Documento de Comunicación*, si bien se incluyen ejemplos ilustrativos de diagramas UML para representar diversos objetos de información. Una visión en profundidad del GSIM supondría una descripción detallada de cada uno de los objetos de información incluidos en el GSIM así como su representación mediante diagramas de clase en UML y las relaciones entre estos objetos.

Directores y Gerentes de Rango Superior	Folletos del GSIM <sup>1</sup>
Gerentes de Rango Medio Estadísticos Expertos en Materias de Análisis Metodólogos	Folletos del GSIM Documento de Comunicación del GSIM
Arquitectos Analistas de negocio Especialistas en metadatos	Documento de Comunicación del GSIM GSIM Clicklable (UML y Glosario)
Arquitecto de soluciones <sup>2</sup>	GSIM Clicklable (UML y Glosario) Fichero de Arquitectura de Empresa

Tabla 17.1: GSIM según los grados de especialización y/o responsabilidades.

## 17.2 Introducción al Documento de Comunicación del GSIM

Las distintas organizaciones estadísticas realizan actividades similares aunque con variaciones en los procesos que cada una de ellas diseña y ejecuta. Cada una de estas actividades hace uso de y produce información similar (por ejemplo, todas las organizaciones usan clasificaciones, crean conjuntos de datos y difunden información). Aunque la información utilizada por las organizaciones es en el fondo la misma, todas las organizaciones tienden a describir esta información de forma ligeramente distinta (y a menudo de forma distinta incluso dentro de cada organización). En el pasado no había una forma común de describir esta información. Esto dificulta la comunicación dentro y entre las organizaciones estadísticas y, sin esta comunicación, no es posible llevar a cabo colaboraciones exhaustivas, estandarización o el intercambio de herramientas y métodos comunes.

El GSIM es el primer marco de referencia avalado internacionalmente para información estadística. Proporciona **un conjunto de objetos de información<sup>3</sup> estandarizados consistentemente descritos**, que aparecen como inputs y outputs en el diseño y producción de estadísticas. Como marco de referencia, el GSIM puede explicar relaciones significativas entre las diversas entidades<sup>4</sup> involucradas en la producción estadística y puede emplearse como guía para el desarrollo y uso de estándares de implementación y especificaciones consistentes.

Como lenguaje común para describir la información estadística, el GSIM puede facilitar la comunicación dentro de y entre organizaciones estadísticas. Puede proporcionar la base para la colaboración, estandarización y compartición de herramientas y métodos y, por tanto, jugar un papel importante en modernizar, mejorar y alinear estándares y producción asociados con las estadísticas oficiales tanto a nivel nacional como internacional.

<sup>3</sup>Traducimos *information object* en el marco de la modelización de la información como *objeto de información*.

<sup>4</sup>Traducimos *entity* en el contexto de la modelización de la información como *entidad*.

El GSIM es uno de los pilares para modernizar la producción estadística oficial alejándose de los compartimentos estancos producidos en cada materia y dominio estadístico. Es un elemento clave para la visión estratégica del *High-Level Group for the Modernisation of Official Statistics* (UNECE 2021c) y está respaldado por la *Conference of European Statisticians* (CES) (UNECE 2021b).

La modernización de la producción estadística es necesaria con el fin de que las organizaciones estadísticas sigan siendo relevantes y flexibles en un entorno informático dinámico y competitivo. Se espera que las organizaciones estadísticas adoptarán e implementarán el GSIM y el lenguaje común que proporciona. Sin embargo, un modelo sólo no puede transformar una organización o sus procesos. Con el fin de alcanzar las necesidades futuras de las organizaciones estadísticas, el GSIM se ha diseñado de forma que permita enfoques innovadores a la producción estadística en la mayor medida posible. Es una de las bases de la *Common Statistical Production Architecture* (CSPA) (UNECE 2021a), una iniciativa colaborativa para diseñar unos servicios comunes e intercambiables con interfaces estándares para ayudar en la estandarización y en la modernización. Al mismo tiempo, el GSIM respalda las formas actuales de producir estadísticas.

### 17.3 Alcance

El GSIM proporciona el marco de objetos de información que sustenta los procesos de producción estadística, tales como los descritos en el *Generic Statistical Business Process Model* (GSBPM) (UNECE 2019c), dando a los objetos de información nombres acordados, definiéndolos, especificando sus propiedades esenciales e indicando sus relaciones con otros objetos de información. No hace, sin embargo, suposiciones sobre los estándares o las tecnologías usadas para implementar el modelo.

El GSIM no incluye objetos de información relacionados con actividades no estadísticas dentro de una organización como la gestión de recursos humanos, económicos o funciones legales, excepto en la medida en que esta información se usa directamente en la producción estadística. Estas actividades se describen en el *Generic Activity Model for Statistical Organisations* (GAMSO) (UNECE 2019a).

El GSIM es un modelo conceptual y no prescribe cómo debe implementarse la información. Las organizaciones pueden escoger estándares que ya existen (como SDMX (SDMX 2012) o DDI (DDI Alliance 2021)) para las implementaciones técnicas.

### 17.4 ¿Qué es el GSIM?

El GSIM contiene objetos que especifican información sobre el mundo real - “objetos de información”. Ejemplos incluyen datos y metadatos (como las clasificaciones esta-



dísticas) así como las reglas y los parámetros necesarios para ejecutar los procesos de producción (por ejemplo, las reglas de depuración de datos). El GSIM identifica en torno a 130 objetos de información, que se agrupan en cinco grupos de alto nivel (véase la Figura 17.1).

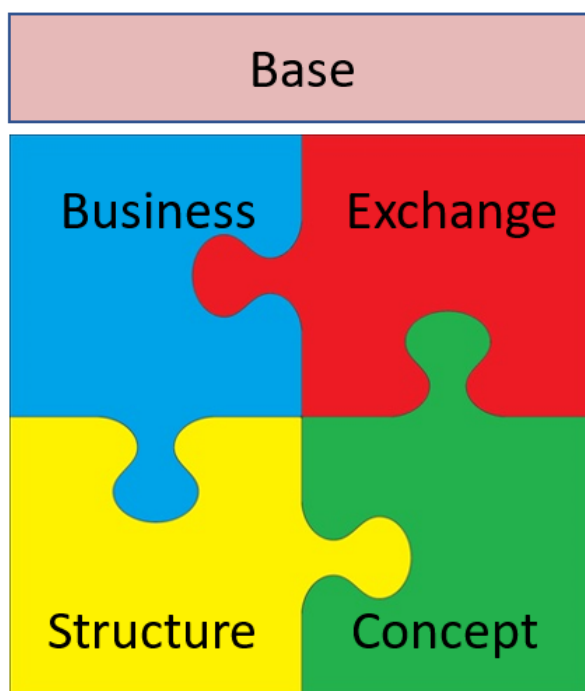


Figura 17.1: Grupos de objetos de información de alto nivel del GSIM.

Los cinco grupos de alto nivel se describen a continuación<sup>5</sup>:

- El Grupo Base<sup>6</sup> proporciona características que se reusan en otros objetos de información para proporcionar funcionalidad como la identificación o versionado de objetos;
- El Grupo de Negocio<sup>7</sup> se usa para expresar los diseños y planes de las *Operaciones Estadísticas*<sup>8</sup> y los procesos que se ejecutan para llevar a cabo tales operaciones estadísticas. Esto incluye la identificación de una *Necesidad Estadística*<sup>9</sup>, los Procesos de Negocio<sup>10</sup> que comprenden la *Operación Estadística* y su *Evaluación*<sup>11</sup>;
- El Grupo de Intercambio<sup>12</sup> se usa para catalogar la información que entra y sale

<sup>5</sup>En cursiva se denotan los objetos de información del GSIM.

<sup>6</sup>Base Group.

<sup>7</sup>Business Group.

<sup>8</sup>Statistical Programs.

<sup>9</sup>Statistical Need.

<sup>10</sup>Business Processes.

<sup>11</sup>Assessment.

<sup>12</sup>Exchange Group.



de una organización estadística a través de los *Canales de Intercambio*<sup>13</sup>. Incluye los objetos de información que describen la recogida y la difusión de la información;

- El Grupo de Concepto<sup>14</sup> se usa para definir el significado de los datos, proporcionando una interpretación de lo que los datos están midiendo;
- El Grupo de Estructuras<sup>15</sup> se usa para estructurar la información a lo largo del proceso de negocio estadístico.

La Figura 17.2 muestra una vista simplificada de los objetos de información del GSIM proporcionando algunos ejemplos de tales objetos en cada grupo.

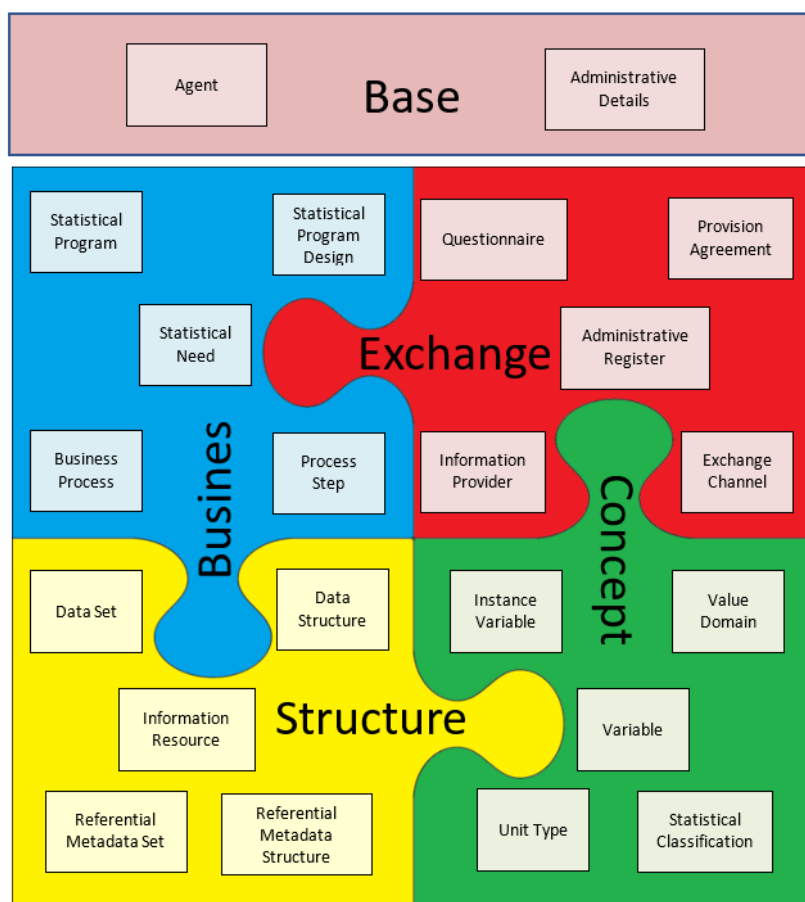


Figura 17.2: Visión simplificada de los objetos de información del GSIM.

La Figura 17.3 muestra otra visión de una parte del GSIM. Ésta es una visión desde un punto de vista más técnico, pero aún así pensada para ser accesible para una audiencia relativamente amplia. Tanto la Figura 17.2 como la Figura 17.3 se pueden usar como forma de comunicación con los usuarios que están interesados en ejemplos de los objetos y de las relaciones en el GSIM.

<sup>13</sup>Exchange Channels.

<sup>14</sup>Concept Group.

<sup>15</sup>Structure Group

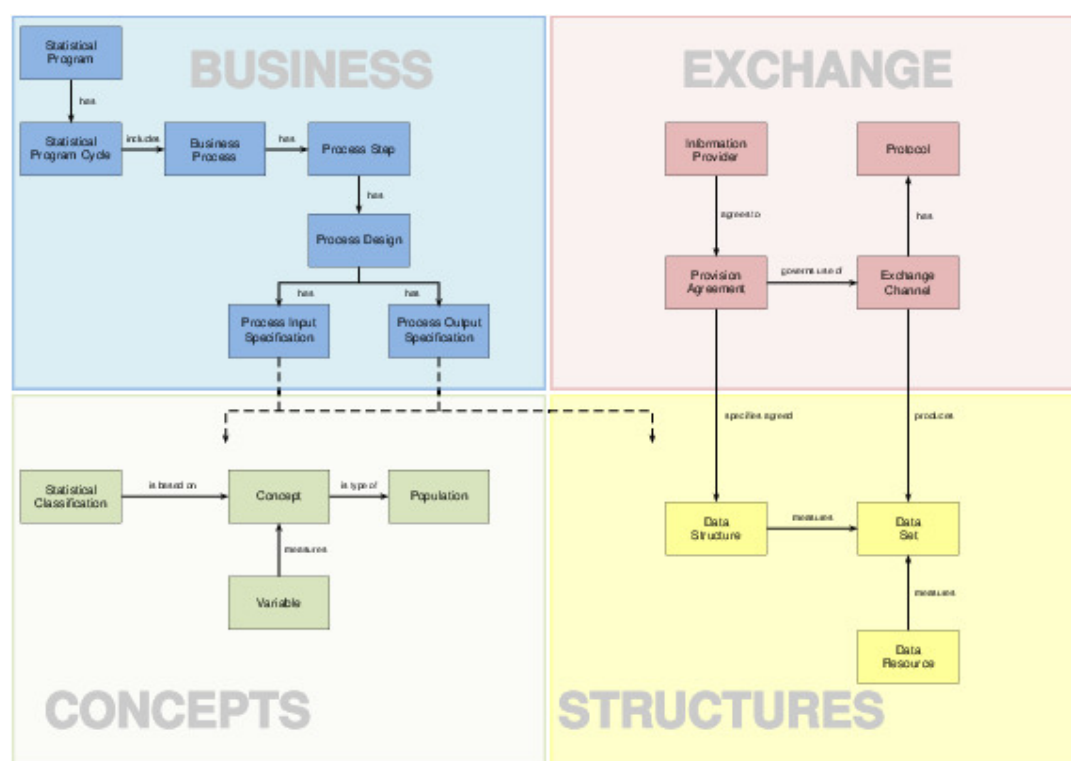


Figura 17.3: Visión alternativa simplificada de los objetos de información del GSIM.

La Figura 17.3 da un ejemplo de objetos de información del GSIM que cuentan una historia sobre la información que es importante en una organización estadística.

Una organización estadística inicia una *Operación Estadística*. La *Operación Estadística* se corresponde con una actividad en curso como una encuesta o una serie de resultados y tiene un *Ciclo de Operación Estadística*<sup>16</sup> (por ejemplo, se repite trimestralmente o anualmente).

El *Ciclo de Operación Estadística* incluirá un conjunto de *Procesos de Negocio*. Los *Procesos de Negocio* consisten en múltiples *Pasos de Proceso*<sup>17</sup> que se especifican en un *Diseño de Proceso*<sup>18</sup>. Estos *Diseños de Proceso* tienen *Especificaciones de Entrada del Proceso*<sup>19</sup> y *Especificaciones de Salida del Proceso*<sup>20</sup>. Las especificaciones a menudo serán elementos de información que se refieren a conceptos y estructuras (por ejemplo, *Clasificación Estadística*<sup>21</sup>, *Variable*,

<sup>16</sup>Statistical Program Cycle.

<sup>17</sup>Process Steps.

<sup>18</sup>Process Design.

<sup>19</sup>Process Input Specifications.

<sup>20</sup>Process Output Specifications.

<sup>21</sup>Statistical Classification.

*Población*<sup>22</sup>, *Estructura de Datos*<sup>23</sup> y *Conjunto de Datos*<sup>24</sup>).

Si, por ejemplo, el *Proceso de Negocio* está relacionado con la recogida de datos, habrá una *Proveedor de Información*<sup>25</sup> que acuerda proporcionar datos a la organización estadística (mediante un *Acuerdo de Provisión*<sup>26</sup>). Este *Acuerdo de Provisión* especifica una *Estructura de Datos* acordada y rige el *Canal de Intercambio*<sup>27</sup> usado para la información entrante. El *Canal de Intercambio* puede ser un *Cuestionario*<sup>28</sup> o un *Registro Administrativo*<sup>29</sup>. Recibirá la información mediante un mecanismo particular (*Protocolo*<sup>30</sup>) como una entrevista o un intercambio de fichero de datos.

El *Conjunto de Datos* producido a través del *Canal de Intercambio* se almacenará en un *Recurso de Datos*<sup>31</sup> y estructurada de acuerdo con una *Estructura de Datos*.

Información más detallada sobre los grupos y sus objetos de información se encuentran en la wiki del [UNECE \(2019b\)](#).

## 17.5 Beneficios del GSIM para la organización como un todo

Se pretende que el GSIM pueda ser utilizado por organizaciones a distintos niveles. Puede ser usado en algunos casos solo como un modelo al que las organizaciones se refieren cuando se comuniquen internamente o con otras organizaciones para clarificar los debates. En otros casos una organización puede elegir implementar el GSIM como el modelo de información que define su entorno operativo. Son válidos varios escenarios para el uso del GSIM, aunque esas organizaciones que hacen uso del GSIM en la mayor medida posible puedan esperar conseguir su máximo beneficio.

### *Beneficios a largo plazo*

El GSIM proporciona un conjunto de objetos de información estandarizados, que son los inputs y los outputs en el diseño y la producción de procesos de negocio estadísticos. Al definir objetos comunes a toda la producción estadística, independientemente del objeto y dominio de análisis, el GSIM permite a las organizaciones estadísticas repensar cómo sus procesos se pueden organizar de una manera más eficiente.

---

<sup>22</sup>Population.

<sup>23</sup>Data Structure.

<sup>24</sup>Data Set.

<sup>25</sup>Information Provider.

<sup>26</sup>Provision Agreement.

<sup>27</sup>Exchange Channel.

<sup>28</sup>Questionnaire.

<sup>29</sup>Administrative Register.

<sup>30</sup>Protocol.

<sup>31</sup>Data Resource.

El GSIM podría ser usado para dirigir la inversión futura hacia áreas de la producción estadística en las que la necesidad común es máxima. También podría permitir cierto grado de especialización dentro de la comunidad estadística internacional. Por ejemplo, algunas organizaciones podrían especializarse en el ajuste estacional, en el análisis de series temporales o en la validación de los datos y otras organizaciones podrían beneficiarse de esta experiencia y pericia.

La implementación del GSIM en combinación con el GSBPM conducirá a ventajas más importantes. El GSIM podría:

- crear un entorno preparado para **reutilizar y compartir métodos, componentes y procesos**;
- proporcionar la oportunidad de implementar un proceso de control basado en reglas y de esta forma minimizar la intervención humana en el proceso de producción;
- facilitar la generación de economías de escala mediante el desarrollo de herramientas comunes de la comunidad de organizaciones estadísticas (un ejemplo muy notable en esta dirección puede encontrarse en [ten Bosch y et al 2021](#)).

#### *Beneficios inmediatos*

Un beneficio significativo de usar el GSIM es que proporciona un **lenguaje común que mejora la comunicación a distintos niveles**:

- entre los distintos roles en los procesos de negocio estadísticos (expertos en las encuestas, metodólogos e informáticos);
- entre los distintos dominios de materias estadísticas;
- entre las diversas organizaciones estadísticas a nivel nacional e internacional.

La mejora de la comunicación dará como resultado un intercambio de datos y metadatos más eficiente dentro de y entre las organizaciones estadísticas y también con los usuarios externos y con los proveedores de datos.

El GSIM puede ser usado ahora por las organizaciones para:

- aumentar la competencia del personal al usar el GSIM como una herramienta de ayuda que proporciona una forma sencilla de entender una visión de la información compleja;
- validar los sistemas de información existentes, comparar buenas prácticas internacionales emergentes y, donde sea apropiado, beneficiarse de experiencias internacionales;
- guiar el desarrollo o actualizar estándares nacionales o internacionales para asegurar que alcanzan las necesidades más amplias de la comunidad estadística internacional.

## 17.6 Relación con otros modelos ModernStats: GSIM y GSBPM

El GSIM presenta vínculos con varios modelos desarrollados bajo los auspicios del *High-Level Group for the Modernisation of Official Statistics* (UNECE 2021c) para dar soporte a la modernización de la producción estadística oficial<sup>32</sup>.

El GSBPM es un modelo de proceso de negocio estadístico que describe y define de manera genérica el conjunto de tareas de producción necesarias para elaborar estadísticas oficiales. Proporciona un marco estándar y terminología armonizada para ayudar a las organizaciones estadísticas a modernizar sus procesos de producción estadística, así como a compartir métodos y componentes del proceso. El GSBPM también se puede usar para integrar estándares de datos y de metadatos, como una plantilla para documentar el proceso, para armonizar infraestructuras informáticas estadísticas y para proporcionar un marco para la evaluación y mejora de la calidad del proceso.

El GSIM y el GSBPM son modelos complementarios para la producción y la gestión de la información estadística. El GSBPM modeliza el proceso de producción estadística e identifica las actividades realizadas por productores de estadísticas oficiales y que da lugar a los outputs de información. Estas actividades están desagregadas en subprocesos como “Depurar e imputar” y “Calcular agregados”. Como se muestra en la Figura 17.4, el GSIM ayuda a describir los subprocesos del GSBPM definiendo los objetos de información que fluyen entre ellos, que son creados en ellos y que son usados por ellos en el proceso de producción de estadísticas oficiales.



Figura 17.4: Relación entre GSIM y GSBPM.

Se obtendrá un mayor valor del GSIM si se aplica conjuntamente con el GSBPM. Sin embargo, es posible (aunque no ideal) aplicar el uno sin el otro. De la misma forma que los procesos estadísticos individuales no usan todos los subprocesos descritos en el GSBPM, es improbable que todos los objetos de información del GSIM sean necesarios en un proceso de negocio estadístico específico.

Una buena gestión de los metadatos es esencial para un buen funcionamiento de los procesos de negocio estadísticos. Los metadatos están presentes en cada fase del GSBPM, creados, actualizados o sin modificar de una fase anterior. En el contexto del GSBPM, el énfasis del proceso generalizado<sup>33</sup> de la gestión de metadatos está en la creación, actualización, uso y reutilización de los metadatos. Las estrategias y sistemas de gestión de los metadatos son, por tanto, vitales para el funcionamiento del GSBPM y son facilitadas

<sup>32</sup>GSBPM (UNECE 2019c), GSIM (UNECE 2019b), GAMSO (UNECE 2019a) y CSPA (UNECE 2021a) son referidos colectivamente como modelos ModernStats.

<sup>33</sup>Overarching process.

por el GSIM.

El GSIM puede asimismo dar soporte a un enfoque consistente a los metadatos facilitando el papel primario de los metadatos de acuerdo con el *Common Metadata Framework* (UNECE 2021d), esto es, los metadatos deben definir de manera única y formal el contenido y los vínculos entre los objetos y procesos del sistema de información estadística.

La descripción de los procesos de negocio estadísticos y sus inputs y outputs mediante el vocabulario estandarizado del GSBPM y del GSIM da cabida a funciones de gestión de la arquitectura (de procesos y, en general, de producción) del siguiente modo:

- facilitando la construcción de sistemas eficientes de recogida, procesamiento y difusión basados en metadatos;
- armonizando las infraestructuras de computación estadísticas;
- diseñando métodos y funciones estandarizados para aplicaciones y herramientas informáticas que dan soporte a los procesos de negocio estadísticos.

Una versión más desglosada de la Figura 17.4 se representa en la Figura 17.5, donde se observan tanto el flujo de datos como el flujo de metadatos<sup>34</sup>.

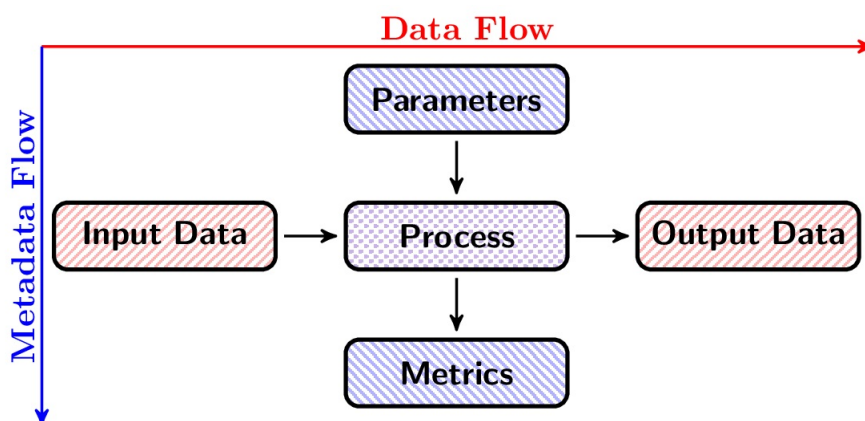


Figura 17.5: Relación entre GSIM y GSBPM en un paso de producción estándar desglosado.

<sup>34</sup>Esta figura está tomada de una conferencia de M. van der Loo en el INE en el año 2018 (véase Loo 2021, para sus referencias).

La CSPA hace uso del GSIM como una referencia común para la definición de información de entrada, de salida y los procesos de negocio. Usar el GSIM como un lenguaje común incrementa la habilidad para comparar información dentro de y entre las organizaciones estadísticas y, por tanto, facilita el desarrollo de servicio y componentes armonizados y reutilizables.

## 17.7 ¿Qué implica para el estadístico?

### *Punto de vista de negocio*

El GSIM ayudará al estadístico a **gestionar de modo efectivo la arquitectura de negocio de una organización estadística** proporcionando una lista estandarizada de los objetos de información empleados en la producción estadística. Establecer las correspondencias reales entre los inputs y outputs de la producción estadística con los objetos de información del GSIM favorece la estandarización transversalmente en los diversos dominios estadísticos de análisis.

El GSIM ayudará al estadístico a mejorar la comunicación con los colegas tanto a nivel local como internacionalmente. La comunicación de temas de análisis entre dominios estadísticos diferentes a menudo es pobre, creando dificultades a la hora de compartir conceptos, variables y componentes de diseño si no se lleva a cabo un ejercicio de correspondencia complejo. El GSIM puede servir como lenguaje común y facilitará la comunicación entre:

- expertos en temas de análisis, metodólogos e informáticos;
- estadísticos de diferentes dominios en una organización estadística;
- estadísticos de diferentes organizaciones.

El GSIM ayudará al estadístico a diseñar y entender mejor los procesos (así como sus inputs y outputs).

Para un ciclo de producción, un estadístico puede diseñar el input, el output y el proceso entre ellos. En términos del GSIM, el input y el output pueden diseñarse en términos de objetos de información de estructuras y conceptos y los procesos intermedios pueden diseñarse usando estos objetos de información. Los objetos de estructuras y conceptos son proporcionados por los expertos y especialistas en los temas de análisis.

Como puede observarse en la Figura 17.6, si se considera el GSBPM como el marco de referencia para los procesos de producción estadística, el primer nivel se puede considerar como equivalente al proceso de producción estadística en su conjunto. El siguiente nivel se corresponde con cada una de las fases del proceso de producción estadístico (por ejemplo, la fase “Procesar” del GSBPM). El tercer nivel se corresponde con los subprocesos (por ejemplo, el subproceso 5.3 del GSBPM - Revisar y validar). El cuarto nivel consiste en los componentes básicos<sup>35</sup> dentro de los subprocesos (como,

---

<sup>35</sup>Traducimos *building blocks* como *componentes básicos*, pero la idea de ladrillos fundamentales a partir



p.ej., la detección de valores monetarios que podrían venir expresados en miles en lugar de unidades).

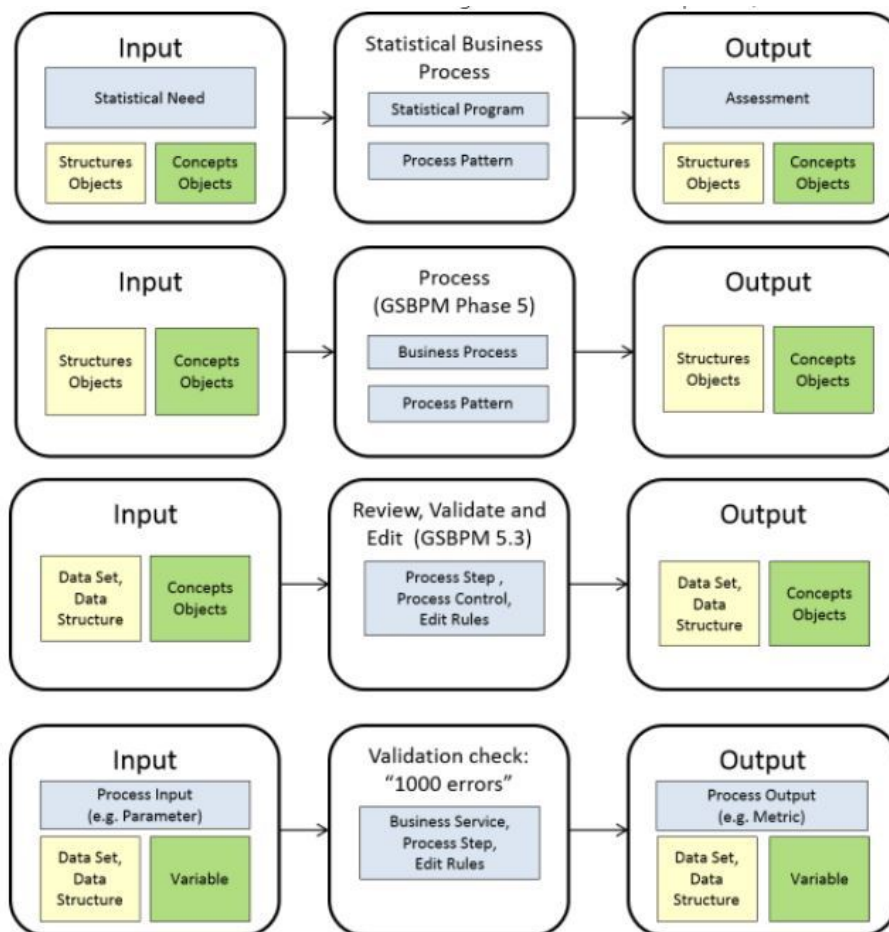


Figura 17.6: Los objetos de información del GSIM en el contexto del GSBPM.

Una cuestión importante para los estadísticos es el problema de componentes de diseño de un único uso, que a menudo se recrean o, por lo menos, son modificadas en cada ciclo de producción. El GSIM facilita la descripción de los inputs y outputs a cada nivel del GSBPM, siguiendo el mismo patrón y así proporcionando una estructura consistente para diseñar los procesos estadísticos. Da soporte al diseño, especificación e implementación de métodos armonizados y de tecnología estándar para crear un sistema de producción estadística generalizado.

El uso del GSIM permitirá la producción de **componentes básicos de proceso reutilizables y flexibles** que pueden ser utilizados por estadísticos para producir productos finales de distinta complejidad, facilitando la producción de una mayor variedad de productos y respondiendo más fácilmente a las necesidades cambiantes de los usuarios.

de los que se crean estructuras más complejas es una imagen muy descriptiva de las ideas subyacentes.



El uso del GSIM, en combinación con otros modelos ModernStats, reducirá las cargas de trabajo ya que muchos procesos pueden ser reconvertidos y reutilizados. Esto significa emplear menos tiempo en trabajos repetitivos y más tiempo para innovación.

A largo plazo, el GSIM, en combinación con otros modelos ModernStats, hará que los estadísticos dependan menos de informáticos y tecnólogos debido a las herramientas diseñadas y desarrolladas para ser parametrizadas para emplearse en proyectos de diferentes dominios.

Los estadísticos están muy interesados en la aplicabilidad, usabilidad y estabilidad de sus métodos y soluciones técnicas. En el enfoque de producción estadística llamado “stove-pipe”<sup>36</sup>, los expertos en el tema de análisis y estadísticos son muy dependientes de los informáticos y tecnólogos para el diseño, construcción y producción de sistemas estadísticos.

Los estadísticos ganarán así mayor control sobre el diseño de sus procesos haciéndolos más autónomos tanto en el diseño como en la producción de sus estadísticas.

La producción se basará sobre aplicaciones más estandarizadas, que son más robustas al cambio y menos vulnerables al cambio de personal. Un aumento en el uso de aplicaciones estandarizadas, que pueden ser fácilmente compartidas entre diversos dominios estadísticos, permitirá a los estadísticos un trabajo más fácil en distintos dominios.

### *Punto de vista de la tecnología de la información*

Una de las principales preocupaciones de los informáticos es la duplicación de esfuerzos debido a la organización “stove-pipe” de la producción estadística. Necesidades poco estables y diversas de las diferentes operaciones estadísticas conducen a soluciones específicas, individuales y a medida, por lo que una gran parte del personal informático desarrolla aplicaciones no estándares y poco documentadas.

La introducción del GSIM tanto a nivel nacional como internacional puede traer beneficios a corto plazo para los informáticos. El GSIM proporciona un lenguaje común para que los informáticos y tecnólogos puedan interactuar con colegas y usuarios internos y externos tanto a nivel local como internacional.

A nivel nacional, los estadísticos se volverán más autónomos en el diseño (ver Figura 17.7) y en la producción de sus estadísticas, reutilizando y readaptando componentes armonizadas ya que el GSIM, en combinación con modelos ModernStats, **permitirá sistemas de producción más flexibles y modulares**. La producción se basará en aplicaciones más estandarizadas que son más robustas al cambio y menos vulnerables a los

---

<sup>36</sup>Enfoque en el que se procuran soluciones individuales para cada operación estadística.

cambios de personal informático. Un aumento del uso de aplicaciones estandarizadas, que pueden ser fácilmente compartidas entre dominios, facilitará el trabajo a los informáticos en los distintos dominios.

El uso del GSIM reducirá la carga de trabajo, ya que muchas componentes pueden ser reconvertidas y reutilizadas. Esto implica menos trabajo repetitivo y más tiempo para innovación.

Esto liberará personal informático para hacer aplicaciones más robustas y explorar nuevos caminos para satisfacer mejor las necesidades cambiantes de las organizaciones estadísticas y sus usuarios. Esto incluirá más tiempo para la creación de procesos robustos, modulares, armonizados y bien documentados que cumplen con las especificaciones de la CSPA.

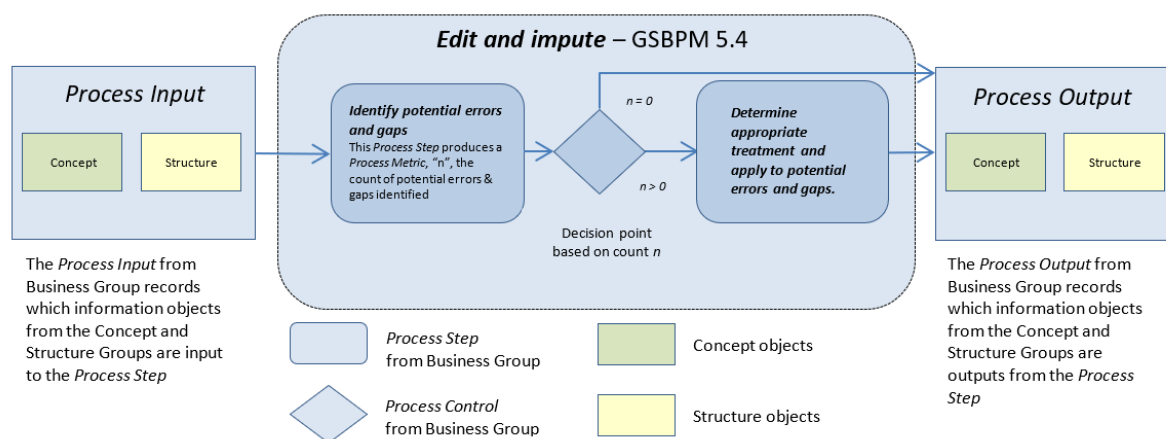


Figura 17.7: Diseño de un proceso de imputación.

A nivel internacional habrá más posibilidades de co-diseñar y co-desarrollar componentes comunes basadas en especificaciones de usuario más robustas de una comunidad de usuarios más amplia. Los desarrolladores informáticos también tendrán acceso a una comunidad más grande en la que todos hablan el mismo idioma para describir su información estadística.

### *Punto de vista de la gestión*

Usar el GSIM (especialmente cuando se combina con el GSBPM) como base para los objetos de información estandarizados puede dar soporte a varias actividades de gestión recogidas en las áreas de actividad del GAMSO (UNECE 2019a):

- para la alta dirección: tomando decisiones y planificando sobre las operaciones estadísticas y las actividades de control;
- para la gestión financiera: controlando el sistema para el proceso de negocio estadístico calculando los costes asociados a los diversos productos estadísticos;

- para la gestión de la calidad: diseñando indicadores de calidad, implementando un marco de calidad y monitorizando la calidad del proceso y de los productos;
- para la gestión de la metodología: diseñando, estandarizando y manteniendo las metodologías estadísticas;
- para la gestión de la información: gestionando el sistema de gestión de datos y metadatos y compilando la estrategia de gestión de los datos.

## 17.8 SDMX, DDI y otros estándares

Como marco de referencia de objetos de información, el GSIM tiene una relación complementaria con estándares como el SDMX ([SDMX 2012](#)) y la DDI ([DDI Alliance 2021](#)), que se usan comúnmente para representar e intercambiar los datos y metadatos estadísticos.

El estándar SDMX proporciona una terminología estándar para datos y metadatos estadísticos, así como estándares técnicos y directrices orientadas hacia el contenido para la transferencia de datos y metadatos, que se podrían aplicar entre subprocesos dentro de una organización estadística. También se considera que el SDMX puede proporcionar el formato para la transmisión de datos entre subprocesos dentro de una organización estadística.

El segundo modelo se ha desarrollado en el consorcio *Data Documentation Initiative*, una iniciativa internacional para establecer un estándar de documentación técnica para describir los datos en ciencias sociales. La Alianza DDI incluye principalmente instituciones académicas y de investigación, por tanto, el alcance de este modelo es un poco distinto del GSIM, que se aplica de manera específica a las organizaciones productoras de estadísticas oficiales. A pesar de esto, el proceso estadístico parece bastante similar entre los productores de estadísticas oficiales y no oficiales, como puede observarse en la consistencia a alto nivel entre los modelos (véase la Figura 17.8).

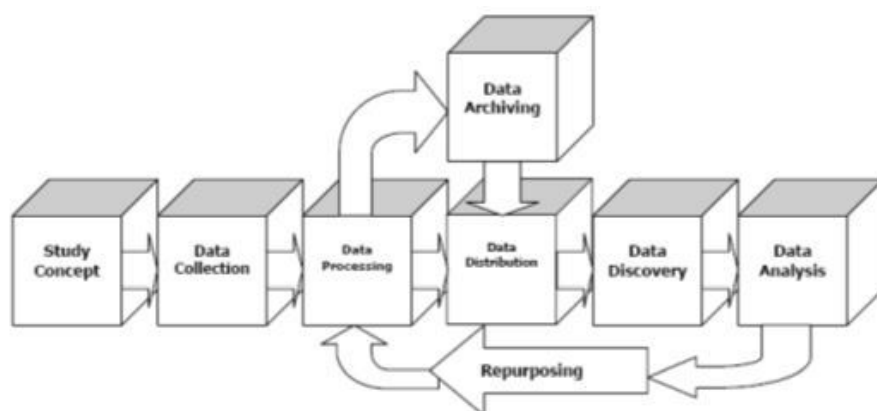


Figura 17.8: El modelo combinado del ciclo de vida DDI 3

Los objetos de información del GSIM son conceptuales; no se estipulan representaciones físicas específicas de la información. Como ejemplo extremadamente simplificado, el

nombre de una organización se puede definir como el mismo concepto independientemente de si la información se graba en una base de datos, en una hoja de cálculo, en un fichero CSV, en un fichero XML o en un trozo de papel.

EL GSIM permite a las organizaciones empezar con un lenguaje común relacionado con los datos y los metadatos utilizados a lo largo de todo el proceso de producción estadística. En este contexto, se han establecido correspondencias entre los objetos de información del GSIM y las representaciones relevantes en el SDMX y/o la DDI.

Esto ayudará a las organizaciones estadísticas a describir y gestionar información estadística usando un lenguaje común mientras, a nivel de los sistemas, la información es representada e intercambiada en un formato apropiado y técnicamente estándar.

Así como los objetos de información del GSIM pueden hacerse corresponder con el SDMX y la DDI (y se pueden obtener beneficios considerables de usar estos estándares), el GSIM no requiere que se utilicen estos estándares. Algunos productores y usuarios de estadísticas pueden decidir usar estándares alternativos por motivos particulares. En otros casos, los productores pueden estar abiertos a usar el SDMX y/o DDI pero tienen sistemas informáticos heredados que no son baratos de actualizar para ser usados con estos estándares.

Describir la información estadística usando el GSIM como un punto común de referencia ayuda a los usuarios a identificar la relación entre dos conjuntos de información estadística que son representados de forma distinta desde una perspectiva técnica.

Por ejemplo, un estadístico puede recibir algunos datos descritos en DDI y algunos descritos con un formato creado localmente. El estadístico puede relacionar ambos con el GSIM. El estadístico será capaz de identificar qué diferencias son puramente técnicas y cuáles reflejan diferencias conceptuales subyacentes.

Una vez que la naturaleza y el alcance de las diferencias son entendidos, a menudo resulta sencillo transformar la información en una representación técnica común (por ejemplo, SDMX o DDI) que permite que el contenido sea integrado y explorado. Este enfoque asegura que los resultados de la conversión técnica a un estándar común se entienden de forma precisa y son adecuados desde una perspectiva conceptual.

Hay un número de sinergias entre el uso del GSIM como marco de referencia y la aplicación de estándares de representación como el SDMX y la DDI. Estas sinergias han sido maximizadas por diseño.

Por ejemplo, para determinar el conjunto de definiciones que se usarán para los objetos de información del GSIM, se utilizaron estándares y modelos existentes como fuentes

de referencia importantes. Así como ninguna de estas fuentes existentes tenía el mismo propósito y alcance que el GSIM - que es un marco de referencia de objetos de información que abarque el proceso de producción estadístico completo - el desarrollo de cada uno implicó el análisis y el soporte para necesidades y escenarios particulares relacionados con tipos particulares de datos y metadatos estadísticos.

De esta forma, el GSIM se benefició de la inversión en tiempo de análisis, modelado, pruebas y perfeccionamiento llevada a cabo cuando se desarrollaron estos estándares y modelos hasta su nivel de madurez actual. Esto también significa que el GSIM no varía “por ninguna razón” de los términos y definiciones que se han usado en estándares y modelos existentes. Donde varía es por razones como la existencia de estándares y modelos relevantes inconsistentes internacionalmente, en los que los estadísticos informaron que términos o definiciones alternativas son más relevantes para sus necesidades.

## Anexos

### Diagramas UML: descripción de la información

El lenguaje UML (véase p.ej. [Booch y col. 2007](#)) es un lenguaje de modelado de sistemas de software que se emplea muy asiduamente para el diseño y análisis de sistemas de información. Se trata de un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. Es importante remarcar que UML es un *lenguaje de modelado* para proporcionar especificaciones de proceso y/o de software y para describir métodos o procesos. No puede considerarse programación (como la programación estructurada). El lenguaje cuenta con varios tipos de diagramas que muestran diferentes aspectos de las entidades representadas.

En el documento de especificaciones del GSIM v1.1 (véase [UNECE 2019b](#), versiones anteriores) pueden encontrarse varios ejemplos de uso de diagramas UML para describir determinados objetos de información relacionados entre sí en el proceso de producción estadística. Se recogen en tal versión las siguientes situaciones: (a) identificación y evaluación de *Necesidades Estadísticas*; (b) diseño y gestión de *Operaciones Estadísticas*; (c) diseño de procesos; (d) ejecución de procesos; (e) intercambio de información; (f) recogida de información; (g) procesamiento y análisis de información y (h) difusión de información. Estas son situaciones genéricas comunes a muchas organizaciones estadísticas.

A modo de ejemplo ilustrativo, en la Figura 17.9 se esquematiza el diseño y gestión de *Operaciones Estadísticas*. Esta figura puede ser descrita con detalle como sigue.

Una organización estadística dará respuesta a una *Necesita Estadística*<sup>37</sup> creando un *Caso*

---

<sup>37</sup>Statistical Need.

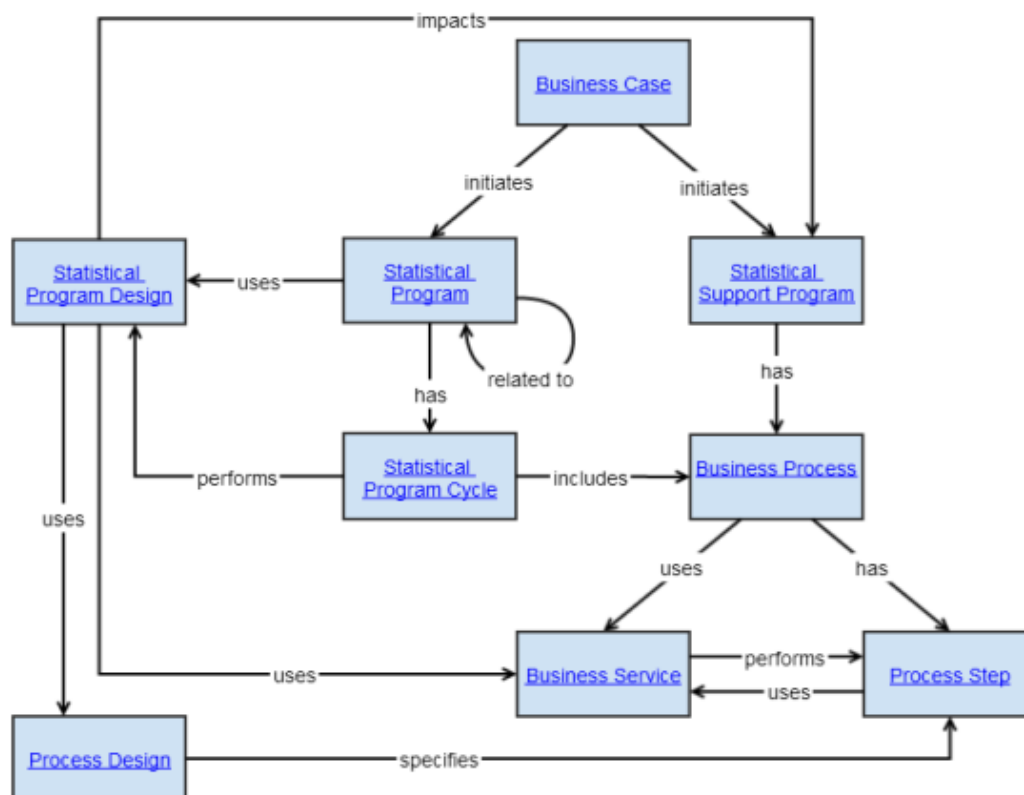


Figura 17.9: Diseño y gestión de Operaciones Estadísticas.

de Negocio<sup>38</sup>. Como consecuencia se producirá una de estas tres circunstancias: (i) la creación de una nueva *Operación de Soporte Estadístico*<sup>39</sup>; (ii) la creación de una nueva *Operación Estadística* o (iii) la evolución de un *Diseño de Operación Estadística*<sup>40</sup> existente que será implementado en una *Operación Estadística* existente.

Las *Operaciones de Soporte Estadístico* conllevan actividades de la organización estadística tales como programas de cambio estadístico, programas de gestión de datos, programas de gestión de metadatos, programas de investigación metodológica, etc. Un buen ejemplo es un programa que gestiona las clasificaciones.

Las *Operaciones Estadísticas* son aquellos programas que una organización lleva a cabo para producir estadísticas (por ejemplo, una encuesta de comercio al por menor). Las *Operaciones Estadísticas* son a menudo cíclicas, esto es, ejecutan ciclos de recogida, producción y difusión de productos estadísticos. Cada uno de estos ciclos se representa mediante un *Ciclo de Operación Estadística*<sup>41</sup>. El *Ciclo de Operación Estadística* es una actividad que se repite para producir estadísticas en períodos de referencia temporal

<sup>38</sup>Business Case.

<sup>39</sup>Statistical Support Program.

<sup>40</sup>Statistical Program Design.

<sup>41</sup>Statistical Program Cycle.

concretos (por ejemplo, la encuesta de comercio al por menor para marzo de 2019).

Las *Operaciones Estadísticas* requieren *Diseños de Operaciones Estadísticas* para alcanzar sus objetivos. Estos diseños cubren el diseño de todas las actividades que deben ejecutarse, notablemente al nivel de *Procesos de Negocio*. En un *Ciclo de Operación Estadística* se ejecutarán usualmente varios *Procesos de Negocio*. Estos se corresponden con los procesos y subprocesos del GSBPM (véase la Figura 17.4). Estos *Procesos de Negocio* pueden repetirse dentro de un ciclo. Cada iteración puede estar compuesta de múltiples actividades del mismo tipo o de tipos diferentes. Como ejemplo, en un mismo ciclo, la *Operación Estadística* podría ejecutar tres iteraciones de la recogida y procesamiento de datos, para luego analizar los datos y difundir los *Productos*<sup>42</sup> estadísticos resultantes. Cada una de estas actividades pueden entenderse como un *Proceso de Negocio* por separado.

EL *Diseño de Operación Estadística* especifica cómo se ejecutarán los *Procesos de Negocio*. Esto incluya la reutilización de *Servicios de Negocio*<sup>43</sup> (posiblemente provenientes de fuera de la organización estadística) o mediante el diseño y uso de procesos más tradicionales. En este último caso, se emplearían objetos de *Diseño de Proceso*<sup>44</sup> para especificar los *Pasos de Producción*. (Aunque los *Servicios de Negocio* reutilizables se especifican también mediante *Diseños de Proceso* y *Pasos de Proceso*, ya existen con antelación y no se necesita trabajo de diseño como parte del *Diseño de la Operación Estadística*).

Debe hacerse notar que serán los *Diseños de Operación Estadística* los que especificarán qué *Pasos de Proceso* requerirán *Diseños de Proceso* y qué *Servicios de Negocio* se emplearán, pero no contienen las especificaciones de bajo nivel la ejecución de tales *Pasos de Proceso* y *Servicios de Negocio*. Estas especificaciones se encuentran en el objeto de *Diseño de Proceso*.

## Diagramas UML: información fundamental

Los objetos de información de carácter fundamental también pueden representarse mediante diagramas UML. En el documento de especificaciones del GSIM v1.1 se detallan especialmente los siguientes: (a) *Conceptos*<sup>45</sup>; (b) *Población*<sup>46</sup>; (c) *Nodo y Conjunto de Nodos*; (d) *Clasificación Estadística*<sup>47</sup>; (e) *Variable*<sup>48</sup>; (f) *Variable Representada*<sup>49</sup>; (g) *Variable de instancia*<sup>50</sup>; (h) *Recursos de Información*<sup>51</sup>; (i) *Conjuntos de Datos*<sup>52</sup>; (j) *Estructuras de Datos de*

---

<sup>42</sup>Product.

<sup>43</sup>Business Service.

<sup>44</sup>Process Design.

<sup>45</sup>Concept.

<sup>46</sup>Population.

<sup>47</sup>Statistical Classification

<sup>48</sup>Variable.

<sup>49</sup>Represented Variable.

<sup>50</sup>Instance Variable.

<sup>51</sup>Information Resources.

<sup>52</sup>Data Sets.



Unidad<sup>53</sup> y Estructuras de Datos Dimensionales<sup>54</sup> y (k) Conjuntos de Metadatos de Referencia<sup>55</sup>.

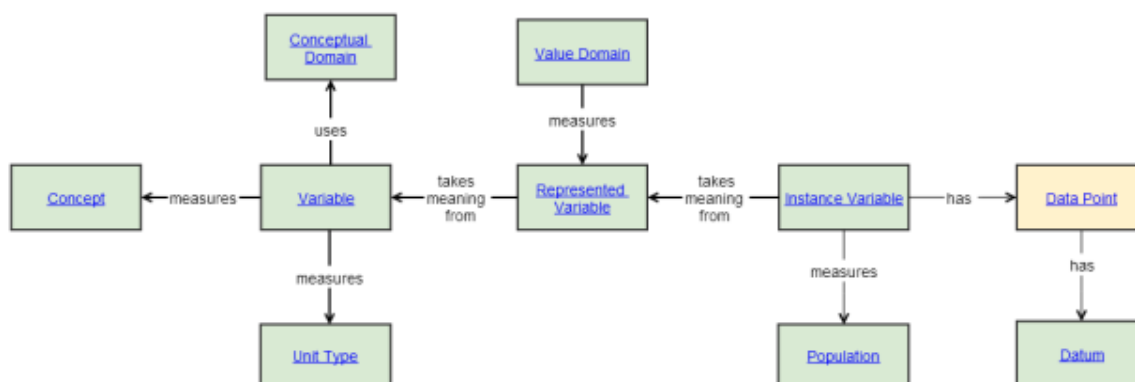


Figura 17.10: Diagrama UML del objeto *Variable de Instancia*.

Como ejemplo ilustrativo incluimos el diagrama UML para especificar el objeto *Variable de Instancia* (véase la Figura 17.10). Una *Variable de Instancia* es una *Variable Representada* que ha sido asociada a un *Conjunto de Datos*. Esto puede corresponder con una columna de datos en un fichero. Por ejemplo, la “edad de todos los residentes mayores de edad en un país en el primer trimestre de 2020” puede ser una columna de datos de una *Variable de Instancia*, que es una combinación de la *Variable Representada* que describe “la edad de tales residentes” y el *Dominio de Valores*<sup>56</sup> de los “números naturales decimales (en años)”.

Un *Dato*<sup>57</sup> está contenido en una *Celda de Datos*<sup>58</sup> en un *Conjunto de Datos*. Puede definirse como la medida de un *Dominio de Valores* asociada con una *Variable de Instancia* combinada con el enlace a una *Unidad* (para *Estructuras de Datos de Unidad*) o a una *Población* (para *Estructuras de Datos Dimensionales*).

## Bibliografía

- Booch, Grady, Robert Maksimchuk, Michael Engle, Bobbi Young, Jim Conallen y Kelli Houston (2007). *Object-Oriented Analysis and Design with Applications, Third Edition*. Third. New York: Addison-Wesley Professional.
- DDI Alliance (2021). *Data Documentation Initiative*. URL: <https://ddialliance.org/>.
- Loo, M. van der (2021). *Home Page*. URL: <http://www.markvanderloo.eu/>.

<sup>53</sup>Unit Data Structure.

<sup>54</sup>Dimensional Data Structure.

<sup>55</sup>Referential Metadata Set.

<sup>56</sup>Value Domain.

<sup>57</sup>Datum.

<sup>58</sup>*Data Point*; escogemos *celda* como traducción de *point* para remarcar la diferencia entre continente y contenido.



- SDMX (2012). *SDMX 2.1 User Guide*. URL: [https://sdmx.org/wp-content/uploads/SDMX\\_2-1\\_User\\_Guide\\_draft\\_0-1.pdf](https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf).
- ten Bosch, O. y et al (2021). *Awesome official statistics software*. Página visitada el día 28 de octubre de 2021. URL: <https://github.com/SNStatComp/awesome-official-statistics-software>.
- UNECE (2019a). *Generic Activity Model for Statistical Organizations*. URL: <https://statswiki.unece.org/display/GAMSO/>.
- (2019b). *Generic Statistical Information Model v1.2*. URL: <https://statswiki.unece.org/display/gsim/>.
  - (2019c). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.
  - (2021a). *Common Statistical Production Architecture*. URL: <https://statswiki.unece.org/display/CSPA>.
  - (2021b). *Conference of European Statisticians*. URL: <https://unece.org/statistics/ces>.
  - (2021c). *High-Level Group for the Modernisation of Official Statistics*. URL: <https://unece.org/statistics/modernization-official-statistics>.
  - (2021d). *The Common Metadata Framework*. URL: <https://statswiki.unece.org/display/hlgbas/The+Common+Metadata+Framework>.

## Tema 18

**La calidad en la estadística oficial y el Código de Buenas Prácticas de las Estadísticas Europeas. El concepto de calidad en la estadística oficial. El Código de Buenas Prácticas de las Estadísticas Europeas. El marco de garantía de la calidad del Sistema Estadístico Europeo. La calidad en los productos y en los procesos estadísticos. Sistemas de evaluación global de la calidad: auditorías, autoevaluación y revisiones por homólogos en las oficinas de Estadística.**

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

- Eurostat (2017). *Código de Buenas Prácticas de las Estadísticas Europeas*. URL: [https://ine.es/ine/codigobp/codigo\\_2017.pdf](https://ine.es/ine/codigobp/codigo_2017.pdf).
- (2019). *El marco de garantía de la calidad del sistema estadístico europeo*. URL: <https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>.
  - (2021). *Peer Review*. URL: <https://ec.europa.eu/eurostat/web/quality/peer-reviews>.
- UNECE (2018). *Indicadores de Calidad del GSBPM*. URL: <https://statswiki.unece.org/display/GSBPM/Quality+Indicators>.
- (2019). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.

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.

## 18.1 El concepto de calidad en la estadística oficial

Las organizaciones que tienen éxito saben que para mantenerse en el negocio la mejora continua es esencial, y que tienen que desarrollar medidas que les ayuden a mejorar. La mejora de la calidad siempre implica un cambio y este debe producirse de acuerdo a un determinado proceso, de la misma forma que para fabricar un coche o para la producción estadística se siguen unos determinados cauces.

Una organización estadística no es diferente al resto de organizaciones en relación con la necesidad de una mejora continua. Existe la necesidad de proporcionar unos resultados con una buena calidad, pero en las organizaciones también hay la necesidad de ser ágiles y ajustar los procesos de acuerdo con las nuevas demandas de los usuarios. En este sentido, ¿cómo se define la calidad?

La calidad se puede definir de forma sencilla como la 'idoneidad de uso'. En el contexto de una encuesta, esto se traduce en la necesidad de que los datos sean tan acurados como sea necesario para conseguir los objetivos establecidos, estar disponibles en el momento en que son necesarios (*timely*, oportunos), y ser accesibles para aquellos para los que se llevó a cabo la encuesta. Precisión, oportunidad (*timeliness*), y accesibilidad, por tanto, son tres dimensiones de la calidad en las encuestas.

Es importante destacar también que la calidad de un producto estadístico es un concepto multidimensional. La calidad de los datos contiene componentes como precisión, oportunidad, riqueza de detalles, accesibilidad, nivel de protección de confidencialidad, etc. Tradicionalmente, el énfasis de la calidad de una encuesta se consideraba en función del error de muestreo (es decir, de la precisión). Sin embargo, como en otras empresas, en las organizaciones estadísticas se ha hecho necesario trabajar en una definición mucho más amplia de calidad, ya que los usuarios no están sólo interesados en la precisión de las estimaciones proporcionadas; a distintos niveles, también es necesario que los datos sean relevantes, oportunos, coherentes, accesibles y comparables.

Algunos han argumentado que la precisión debe de ser siempre lo primero, sin precisión el resto de dimensiones de la calidad son irrelevantes. Sin embargo, lo contrario también es cierto. Datos muy acurados no son útiles si se publican muy tarde o si no son relevantes. La precisión se define normalmente en términos del error de encuesta total (*total survey error*). Sin embargo, algunas fuentes de error son imposibles de medir. En su lugar se debe de *asegurar* la calidad. Y, para ello, el concepto de calidad de Eurostat se basa en las nueve dimensiones que se presentan en la siguiente tabla.

---

Dimensión	Breve descripción
1. Relevancia de un concepto estadístico	Un producto estadístico es relevante si cumple con las necesidades de los usuarios.
2. Precisión de las estimaciones	La precisión es la diferencia entre la estimación y el verdadero valor del parámetro. Evaluar la precisión no siempre es posible.
3. Oportunidad al difundir los resultados	Es quizá una de las más importantes necesidades de los usuarios. Es el tiempo que pasa entre que ocurre el fenómeno que se quiere medir y la disponibilidad de los datos sobre este fenómeno.
4. Puntualidad al difundir los resultados	Tiempo que pasa desde el momento de publicación real de los datos y el momento en que deberían de ser publicados (target).
5. Accesibilidad y claridad de la información	Los resultados son de un gran valor si son fácilmente accesibles y están disponibles en la forma adecuada para los usuarios. También es necesario que los metadatos estén disponibles.
6. Comparabilidad	Comparaciones fiables en el espacio y en el tiempo son a menudo cruciales. Las comparaciones entre países son muy comunes.
7. Coherencia	Cuando se usan distintas fuentes o estudios de distintas periodicidades, las estadísticas son coherentes si están basadas en las mismas definiciones, clasificaciones y estándares metodológicos.
8. Completitud	Los dominios para los cuales las estadísticas están disponibles deberían reflejar las necesidades y prioridades expresadas por los usuarios.
9. Fiabilidad	Cercanía entre el valor inicial estimado y las estimaciones posteriores.

Las circunstancias han cambiado mucho en los últimos años (recortes presupuestarios, nuevas fuentes de datos, nuevos productores de datos), y como consecuencia de esta nueva situación cada vez más organizaciones estadísticas están trabajando con modelos de gestión de la calidad, modelos de excelencia del negocio, orientación de los usuarios, auditorías, y autoevaluaciones como formas de mejorar su trabajo. Esta visión que se ha descrito del proceso en las encuestas alcanza prácticamente a todos los procesos en una organización estadística, ya que todos los procesos intermedios que sirven de apoyo al

trabajo en las encuestas tienen un efecto en la calidad de los productos estadísticos.

## 18.2 El Código de Buenas Prácticas de las Estadísticas Europeas

El *Código de Buenas Prácticas de las Estadísticas Europeas* (CBPEE) ([Eurostat 2017](#)) establece el estándar para desarrollar, producir y difundir las estadísticas europeas. El CBPEE se basa en dieciséis principios, que abarcan el entorno institucional, los procesos estadísticos y la producción estadística. Un conjunto de indicadores de las mejores prácticas y estándares para cada uno de los principios sirve de orientación y referencia para analizar la aplicación del Código, lo que permite aumentar la transparencia dentro del marco del Sistema Estadístico Europeo (SEE).

El CBPEE es un instrumento autorregulador que contiene los estándares para la independencia de las autoridades estadísticas nacionales y europeas y proporciona una garantía del buen funcionamiento de la Sistema Estadístico Europeo y la producción de estadísticas de alta calidad y fiables. La evaluación de su cumplimiento se basa en un mecanismo de *peer reviews*, véase la Sección [18.5](#).

Los fines del CBPEE son:

- Establecer el estándar para el desarrollo, producción y difusión de las estadísticas europeas. Se basa en una definición de calidad de las estadísticas común a todo el SEE y que afecta a todos los ámbitos desde el entorno institucional, a los procesos de producción estadística, y a los productos estadísticos.
- Asegurar la calidad y la credibilidad de los datos. Los principios hacen referencia, entre otros aspectos, a la independencia profesional, la protección de la confidencialidad, la fiabilidad de los resultados, su precisión, oportunidad, puntualidad, accesibilidad, claridad, comparabilidad y coherencia.

El CBPEE va dirigido a:

- Los usuarios: para mostrar que las estadísticas son imparciales, objetivas y fidedignas.
  - Las autoridades encargadas de la gobernanza (Gobiernos, Ministerios): para asegurar que los servicios estadísticos están organizados profesionalmente y dotados de recursos de manera que la independencia, integridad y responsabilidad estén garantizadas.
  - Los proveedores de datos: para demostrar que la confidencialidad de la información está protegida y no supone una carga demasiado grande.
  - Las autoridades estadísticas: para proporcionar una referencia de principios, valores y buenas prácticas que permitan producir y difundir estadísticas de calidad y armonizadas.
-

El Reglamento (CE) N° 223/2009 relativo a la estadística europea <sup>1</sup> crea un nuevo organismo, el *European Statistical Governance Advisory Board* (ESGAB), cuyo objetivo es proporcionar una visión general independiente del SEE en relación con la implementación del CBPEE. La ESGAB asesora a Eurostat sobre las medidas adecuadas que faciliten la implementación del CBPEE, sobre cómo comunicarlo a los usuarios y a los informantes y sobre la actualización del CBPEE.

El CBPEE fue adoptado por la UE en 2005 y fue revisado por el Comité del Sistema Estadístico Europeo en dos ocasiones. La segunda versión se publicó en septiembre de 2011 y la tercera en noviembre de 2017.

La primera versión, consistente en quince principios (los actuales sin el principio 1 bis), tenía dos propósitos:

- Mejorar la confianza de los usuarios proponiendo determinados acuerdos institucionales y organizativos <sup>2</sup>.
- Reforzar la calidad de las estadísticas producidas y difundidas, promoviendo la aplicación coherente de mejores principios, métodos y prácticas estadísticas internacionales por parte de todos los productores de estadísticas oficiales en Europa.

Los principales cambios entre la primera y la segunda versión fueron:

- Se enfatiza el papel coordinador de los Institutos Nacionales de Estadística (INEs) a nivel nacional y el papel y posición de los directores de los mismos. Se deja también claro que el principio de independencia profesional de los INEs se tiene que cumplir incondicionalmente libre de toda presión.
- Fomenta el uso de los registros administrativos clarificando su papel en el diseño del contenido de los registros administrativos y los requisitos de calidad aplicables a los datos administrativos.

Además, el CBPEE se revisó de manera que se distinguiese entre los principios a implementar por parte de los INEs y los principios relativos al entorno institucional, a implementar por parte de los gobiernos de los países miembros.

Las principales novedades introducidas en la revisión de noviembre de 2017 son de tres tipos: por un lado, se ha introducido un principio dedicado a la coordinación y a

---

<sup>1</sup><https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:02009R0223-20150608&from=ES>

<sup>2</sup>Eurostat fue el organismo encargado de verificar los criterios de convergencia para la entrada en la Unión Monetaria. Los datos presentados por Grecia presentaron divergencias con la realidad contable y esto llevó a una crisis de confianza en 2004. En este contexto se elabora la 'Recommendation on the independence, integrity and accountability of the national and Community statistical authorities' <https://ec.europa.eu/eurostat/documents/34693/344802/Commission-recommendation-COM-2005> que es donde se menciona por primera vez la necesidad de elaboración del CBPEE

la cooperación en los sistemas estadísticos; al mismo tiempo, se concede una mayor relevancia a las fuentes estadísticas alternativas a las encuestas, como las fuentes administrativas o las nuevas fuentes de información procedentes de la revolución digital; un tercer ámbito ha sido la inclusión de las novedades acaecidas en el marco normativo comunitario.

A continuación se indican los tres bloques, los dieciséis principios del CBPEE y los indicadores de cada principio:

- **Entorno institucional.** Los factores institucional y organizativo tienen una influencia considerable en la eficacia y en la credibilidad de una autoridad estadística que desarrolla, elabora y difunde estadísticas europeas. Los principios pertinentes son: independencia profesional, coordinación y cooperación, mandato de recogida de datos, adecuación de los recursos, compromiso de calidad, confidencialidad estadística e imparcialidad y objetividad.

– **PRINCIPIO 1. Independencia profesional.** La independencia profesional de las autoridades estadísticas frente a otros departamentos y organismos políticos, reguladores o administrativos, y frente a los operadores del sector privado, garantiza la credibilidad de las estadísticas europeas.

Indicador 1.1 La independencia de los institutos nacionales de estadística y de Eurostat frente a las injerencias externas, tanto políticas como de otro tipo, en el desarrollo, la elaboración y la difusión de estadísticas está establecida por ley y garantizada para otras autoridades estadísticas.

Indicador 1.2 Los directores de los institutos nacionales de estadística y de Eurostat y, en su caso, los directores de otras autoridades estadísticas tienen un nivel jerárquico lo suficientemente elevado como para garantizar un acceso de alto nivel a las autoridades políticas y a los organismos públicos administrativos; son personas de la más alta talla profesional.

Indicador 1.3 Los directores de institutos nacionales de estadística y de Eurostat y, en su caso, los directores de otras autoridades estadísticas tienen la responsabilidad de garantizar que las estadísticas se desarrollen, elaboren y difundan de forma independiente.

Indicador 1.4 Los directores de los institutos nacionales de estadística y de Eurostat y, en su caso, los directores de otras autoridades estadísticas son los únicos responsables para decidir los métodos, las normas y los procedimientos estadísticos, así como el contenido y el calendario de la publicación de las estadísticas.

Indicador 1.5 Los programas de trabajo estadístico se publican y los progresos realizados se describen en informes periódicos.

Indicador 1.6 Las publicaciones de estadísticas se distinguen claramente de las declaraciones políticas y se emiten al margen de estas.

Indicador 1.7 El instituto nacional de estadística, Eurostat y, en su caso, otras autorida-

---

des estadísticas realizan comentarios públicos sobre cuestiones estadísticas, incluyendo críticas y usos inadecuados de las estadísticas cuando procede.

Indicador 1.8 Los procedimientos de contratación y nombramiento de los directores de los institutos nacionales de estadística, de Eurostat y, en su caso, de otras autoridades estadísticas son transparentes y se basan únicamente en la capacidad profesional. Las razones por las que se puede poner fin al ejercicio de un cargo se determinan en el marco legal y no pueden ser razones que comprometan la independencia profesional o científica.

- **PRINCIPIO 1 bis. Coordinación y cooperación.** Los institutos nacionales de estadística y Eurostat garantizan la coordinación de todas las actividades para el desarrollo, la elaboración y la difusión de estadísticas europeas a nivel del sistema estadístico nacional y del Sistema Estadístico Europeo, respectivamente. Las autoridades estadísticas colaboran de manera activa en el marco de cooperación constituido por el Sistema Estadístico Europeo, con el fin de garantizar el desarrollo, la elaboración y la difusión de estadísticas europeas.

Indicador 1bis.1 Los institutos nacionales de estadística coordinan las actividades estadísticas de todas las demás autoridades nacionales que desarrollan, elaboran y difunden estadísticas europeas. En este sentido, actúan como el único punto de contacto para Eurostat en asuntos estadísticos. Cuentan con una legislación y unos procedimientos claros y sólidamente establecidos para implementar la coordinación a nivel nacional y europeo.

Indicador 1bis.2 Cuando es necesario, los directores de los institutos nacionales de estadística imparten orientaciones nacionales, para garantizar la calidad en el desarrollo, la elaboración y la difusión de estadísticas europeas en el marco del sistema estadístico nacional, cuya aplicación es objeto de seguimiento y análisis.

Indicador 1bis.3 Las autoridades estadísticas mantienen y desarrollan de manera continua y a diferentes niveles la cooperación entre sí y con los organismos consultivos del Sistema Estadístico Europeo, así como con los miembros del Sistema Europeo de Bancos Centrales, las instituciones académicas y otros organismos internacionales, según corresponda.

- **PRINCIPIO 2. Mandato de recogida de datos y acceso a los datos.** Las autoridades estadísticas tienen un mandato jurídico claro de recogida de información procedente de varias fuentes de datos, y de acceso a ella, con el fin de elaborar estadísticas europeas. A petición de las autoridades estadísticas, se puede obligar por ley a las administraciones, empresas, hogares y público en general a permitir el acceso a los datos destinados a la elaboración de estadísticas europeas o a facilitar dichos datos.

Indicador 2.1 El mandato de las autoridades estadísticas a la recogida de información procedente de varias fuentes de datos, o el acceso a ella, con el fin de

---



desarrollar, elaborar y difundir estadísticas europeas está establecido por ley.

Indicador 2.2 Las autoridades estadísticas están autorizadas por ley a acceder a los datos administrativos, rápidamente y de forma gratuita, y a utilizarlos con fines estadísticos. Desde el principio participan en el diseño, el desarrollo y la supresión de los registros administrativos, con el fin de adecuarlos en mayor medida a los fines estadísticos.

Indicador 2.3 Sobre la base de un acto jurídico, las autoridades estadísticas pueden obligar a responder a encuestas estadísticas.

Indicador 2.4 Se facilita el acceso con fines estadísticos a otros datos, como los privados, al tiempo que se garantiza la confidencialidad estadística y la protección de los datos.

– **PRINCIPIO 3. Adecuación de los recursos.** Los recursos a disposición de las autoridades estadísticas son suficientes para cumplir los requisitos de las estadísticas europeas.

Indicador 3.1 Se dispone de recursos humanos, financieros y técnicos adecuados, tanto en tamaño como en calidad, para cumplir las necesidades en materia de estadística.

Indicador 3.2 El ámbito, el detalle y el coste de las estadísticas son proporcionados con respecto a las necesidades.

Indicador 3.3 Existen procedimientos para evaluar y justificar las solicitudes de nuevas estadísticas en relación con su coste.

Indicador 3.4 Existen procedimientos para evaluar la necesidad de realizar cada estadística, a fin de determinar si alguna de ellas puede eliminarse o reducirse para liberar recursos.

– **PRINCIPIO 4. Compromiso de calidad.** Las autoridades estadísticas están comprometidas con la calidad; identifican sistemática y regularmente los puntos fuertes y débiles para mejorar continuamente la calidad del proceso y del producto.

Indicador 4.1 Existe una política de calidad definida y está a disposición del público. Se dispone de una estructura y unas herramientas organizativas para gestionar la calidad.

Indicador 4.2 Se han establecido procedimientos para la planificación, el seguimiento y la mejora de la calidad de los procesos estadísticos, incluida la integración de datos procedentes de varias fuentes.

Indicador 4.3 La calidad del producto se controla con regularidad, se evalúa con respecto a posibles compromisos entre componentes de calidad y es objeto de informes acordes con los criterios de calidad europeos.

---

Indicador 4.4 Se lleva a cabo un análisis periódico y minucioso de la producción estadística clave, recurriendo incluso a expertos externos cuando es necesario.

- **PRINCIPIO 5. Confidencialidad estadística y protección de datos.** La privacidad de los informantes, la confidencialidad de la información que proporcionan, su uso exclusivo con fines estadísticos y la seguridad de los datos están totalmente garantizados.

Indicador 5.1 La confidencialidad estadística está garantizada por ley.

Indicador 5.2 El personal firma un compromiso jurídico de confidencialidad cuando es contratado.

Indicador 5.3 Se han establecido sanciones para cualquier incumplimiento deliberado de la confidencialidad estadística.

Indicador 5.4 Se imparten al personal orientaciones e instrucciones sobre la protección de la confidencialidad estadística a lo largo de todos los procesos estadísticos. La política de confidencialidad está a disposición del público.

Indicador 5.5 Se han adoptado las medidas reguladoras, administrativas, técnicas y organizativas necesarias para proteger la seguridad y la integridad de los datos estadísticos y su transmisión, de conformidad con las mejores prácticas, las normas internacionales, así como la legislación europea y nacional.

Indicador 5.6 Se aplican protocolos estrictos a los usuarios externos que acceden a microdatos estadísticos con fines de investigación.

- **PRINCIPIO 6. Imparcialidad y objetividad.** Las autoridades estadísticas desarrollan, elaboran y difunden estadísticas europeas respetando la independencia científica y de forma objetiva, profesional y transparente, de modo que todos los usuarios reciben el mismo trato.

Indicador 6.1 Las estadísticas se recopilan sobre una base objetiva determinada por consideraciones estadísticas.

Indicador 6.2 La elección de las fuentes de datos y los métodos estadísticos, así como las decisiones en cuanto a la difusión de las estadísticas, se basan en consideraciones estadísticas.

Indicador 6.3 Los errores detectados en las estadísticas publicadas se corrigen lo antes posible y dichas correcciones se hacen públicas.

Indicador 6.4 La información sobre las fuentes de datos, los métodos y los procedimientos utilizados están a disposición del público.

Indicador 6.5 Se anuncian con antelación la fecha y la hora de publicación de las estadísticas.

Indicador 6.6 Se anuncian por adelantado las revisiones o cambios importantes en la metodología.

---

Indicador 6.7 Las autoridades estadísticas deciden de manera independiente el momento y el contenido de la publicación de las estadísticas, sin perder de vista el objetivo de suministrar una información estadística completa y oportuna. Todos los usuarios tienen igual acceso y al mismo tiempo a la publicación de las estadísticas. El acceso privilegiado de cualquier usuario externo previo a la publicación tiene carácter limitado, debidamente justificado, controlado y público. Si se producen filtraciones, se revisan los acuerdos de acceso privilegiado a fin de garantizar la imparcialidad.

Indicador 6.8 Las publicaciones y declaraciones estadísticas realizadas en ruedas de prensa son objetivas e imparciales.

- **Procesos estadísticos.** En los procesos estadísticos utilizados por las autoridades estadísticas para desarrollar, elaborar y difundir estadísticas europeas, se respetan plenamente las normas, orientaciones y buenas prácticas, tanto europeas como internacionales, y se hace un esfuerzo constante por buscar la innovación. La credibilidad de las estadísticas se ve reforzada por una reputación de buena gestión y eficacia. Los principios pertinentes son: metodología sólida, procedimientos estadísticos adecuados, carga no excesiva para los encuestados y rentabilidad.

- **PRINCIPIO 7. Metodología sólida.** Las estadísticas de calidad se apoyan en una metodología sólida, que requiere herramientas, procedimientos y conocimientos adecuados.

Indicador 7.1 El marco metodológico general de las estadísticas europeas sigue normas, orientaciones y buenas prácticas tanto europeas como internacionales y hace un esfuerzo constante en busca de la innovación.

Indicador 7.2 Se dispone de procedimientos para garantizar que en todas las instancias de la autoridad estadística se aplican de manera consistente conceptos, definiciones, clasificaciones y otros estándares.

Indicador 7.3 Los registros y marcos utilizados para la elaboración de estadísticas europeas se evalúan periódicamente y, en su caso, se ajustan para garantizar un alto nivel de calidad.

Indicador 7.4 Existe una concordancia detallada entre los sistemas de clasificación nacionales y los sistemas europeos correspondientes.

Indicador 7.5 Se contratan titulados en las disciplinas académicas pertinentes.

Indicador 7.6 Las autoridades estadísticas establecen una política de formación profesional continua para su personal.

Indicador 7.7 Las autoridades estadísticas mantienen y desarrollan la cooperación con la comunidad científica para mejorar la metodología y la eficacia de los métodos aplicados y para promover mejores herramientas cuando es viable.

- **PRINCIPIO 8. Procedimientos estadísticos adecuados.** Las estadísticas de calidad se apoyan en procedimientos estadísticos adecuados, aplicados a lo
-

largo de todos los procesos estadísticos.

Indicador 8.1 Cuando las estadísticas europeas se basan en datos administrativos y de otro tipo, las definiciones y los conceptos utilizados con fines no estadísticos son una buena aproximación a los requeridos para fines estadísticos.

Indicador 8.2 En el caso de las encuestas estadísticas, se prueban sistemáticamente los cuestionarios antes de la recogida de datos.

Indicador 8.3 Los procesos estadísticos se someten a un control rutinario y se revisan según procede.

Indicador 8.4 Los metadatos relacionados con procesos estadísticos se gestionan a lo largo de todos los procesos estadísticos y se difunden según procede.

Indicador 8.5 Las revisiones siguen procedimientos normalizados, consolidados y transparentes.

Indicador 8.6 Se establecen acuerdos con los titulares de los datos administrativos y de otro tipo en los que se plasma el compromiso común de utilizar dichos datos con fines estadísticos.

Indicador 8.7 Las autoridades cooperan con los titulares de los datos administrativos y de otro tipo para garantizar la calidad de los datos.

– **PRINCIPIO 9. Carga no excesiva para los encuestados.** La carga de respuesta es proporcionada a las necesidades de los usuarios y no es excesiva para los encuestados. Las autoridades estadísticas controlan la carga de respuesta y fijan objetivos para reducirla progresivamente.

Indicador 9.1 El alcance y el detalle de las exigencias en materia de estadísticas europeas se limitan a lo estrictamente necesario.

Indicador 9.2 La carga de respuesta se reparte todo lo posible entre las poblaciones objeto de la encuesta y es controlada por la autoridad estadística.

Indicador 9.3 En la medida de lo posible, se puede acceder fácilmente a la información que se solicita de las empresas a partir de sus cuentas y, cuando es posible, se utilizan medios electrónicos para facilitar su transmisión.

Indicador 9.4 Siempre que es posible se utilizan fuentes de datos administrativos y de otro tipo para evitar que se dupliquen las solicitudes de datos.

Indicador 9.5 A fin de minimizar la carga de respuesta, se promueve el uso compartido de datos y su integración, condicionados al cumplimiento de requisitos de confidencialidad y protección de datos.

Indicador 9.6 A fin de minimizar la carga de respuesta, las autoridades estadísticas promueven medidas que permitan enlazar las distintas fuentes de datos.

– **PRINCIPIO 10. Rentabilidad.** Los recursos se utilizan eficientemente.

---

Indicador 10.1 Medidas internas y externas independientes controlan el uso de recursos por parte de la autoridad estadística.

Indicador 10.2 El potencial productivo de las tecnologías de la información y de las comunicaciones se utiliza para optimizar los procesos estadísticos.

Indicador 10.3 Se realizan esfuerzos proactivos por mejorar el potencial estadístico de las fuentes de datos administrativos y de otro tipo y limitar el uso de encuestas directas.

Indicador 10.4 Las autoridades estadísticas promueven, comparten y aplican soluciones normalizadas que mejoran la eficacia y la eficiencia.

- **Producción estadística.** Las estadísticas disponibles satisfacen las necesidades de los usuarios. Las estadísticas cumplen las normas de calidad europeas y responden a las necesidades de las instituciones europeas, los gobiernos, los centros de investigación, las empresas y el público en general. La calidad del producto se mide por el grado de pertinencia, precisión y fiabilidad, oportunidad, coherencia, comparabilidad entre regiones y países, así como de facilidad de acceso para los usuarios, de las estadísticas. Es decir, por los principios de la producción estadística.

- **PRINCIPIO 11. Pertinencia.** Las estadísticas europeas satisfacen las necesidades de los usuarios.

Indicador 11.1 Existen procedimientos para consultar a los usuarios, controlar la pertinencia y el valor de las estadísticas existentes por lo que se refiere a sus necesidades, y para considerar sus nuevas necesidades y prioridades y anticiparse a ellas. Se persigue la innovación, para mejorar continuamente la producción estadística.

Indicador 11.2 Se satisfacen las necesidades prioritarias y se reflejan en el programa de trabajo.

Indicador 11.3 Se realiza un control periódico y un seguimiento sistemático de la satisfacción de los usuarios.

- **PRINCIPIO 12. Precisión y fiabilidad.** Las estadísticas europeas reflejan la realidad de manera precisa y fiable.

Indicador 12.1 Los datos originales, los datos integrados, los resultados intermedios y la producción estadística se evalúan y validan periódicamente.

Indicador 12.2 Los errores de muestreo y los ajenos al muestreo se calculan y se documentan sistemáticamente con arreglo a las normas europeas.

Indicador 12.3 Se analizan periódicamente las revisiones, a fin de mejorar los datos originales, los procesos estadísticos y el producto.

- **PRINCIPIO 13. Oportunidad y puntualidad.** Las estadísticas europeas se publican oportuna y puntualmente.
-

- Indicador 13.1 El calendario de publicación de las estadísticas es conforme con las normas europeas e internacionales al respecto.
- Indicador 13.2 Se hace pública una hora determinada del día para la publicación de estadísticas.
- Indicador 13.3 La periodicidad de las estadísticas tiene en cuenta las necesidades de los usuarios en la medida de lo posible.
- Indicador 13.4 Cuando no se cumple el calendario previsto de difusión, se notifica por adelantado, se dan explicaciones y se establece una nueva fecha de publicación.
- Indicador 13.5 Si se considera útil, pueden hacerse públicos resultados preliminares con una precisión y una fiabilidad agregadas aceptables.

– **PRINCIPIO 14. Coherencia y comparabilidad.** Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de diferentes fuentes de datos.

- Indicador 14.1 Las estadísticas son coherentes y consistentes internamente (es decir, se observan identidades aritméticas y contables).
- Indicador 14.2 Las estadísticas son comparables durante un período de tiempo razonable.
- Indicador 14.3 Las estadísticas se recopilan sobre la base de normas comunes con respecto al alcance, las definiciones, las unidades y las clasificaciones en las distintas encuestas y fuentes de datos.
- Indicador 14.4 Las estadísticas procedentes de distintas fuentes de datos y con distinta periodicidad se comparan y ajustan entre sí.
- Indicador 14.5 La comparabilidad transnacional de los datos dentro del Sistema Estadístico Europeo se garantiza mediante intercambios periódicos entre dicho sistema y otros sistemas estadísticos. Se realizan estudios metodológicos en estrecha colaboración entre los Estados miembros y Eurostat.

– **PRINCIPIO 15. Accesibilidad y claridad.** Las estadísticas europeas se presentan de forma clara y comprensible, se difunden de forma adecuada y conveniente, su disponibilidad y acceso tienen carácter imparcial y van acompañadas de metadatos y orientación de apoyo.

- Indicador 15.1 Las estadísticas y los metadatos correspondientes se presentan y se archivan de tal forma que facilitan la interpretación adecuada y las comparaciones significativas.
- Indicador 15.2 Los servicios de difusión utilizan modernas tecnologías de la información y de las comunicaciones, métodos y plataformas, así como estándares para datos abiertos.
-

- Indicador 15.3 Cuando es posible, se suministran análisis a medida y se informa de ello al público.
- Indicador 15.4 El acceso a los microdatos está permitido con fines de investigación y está sujeto a normas o protocolos específicos.
- Indicador 15.5 Los metadatos relacionados con productos los gestiona y difunde la autoridad estadística de conformidad con las normas europeas.
- Indicador 15.6 Se mantiene informados a los usuarios sobre la metodología de los procesos estadísticos, incluido el uso y la integración de datos administrativos y de otro tipo.
- Indicador 15.7 Se mantiene informados a los usuarios sobre la calidad de la producción estadística con respecto a los criterios de calidad de las estadísticas europeas.

### 18.3 El marco de garantía de la calidad del Sistema Estadístico Europeo

Para llevar a cabo actividades de evaluación en una organización estadística, **los marcos de garantía de la calidad y los marcos institucionales** tienen como objetivo establecer un sistema de **métodos y herramientas coordinados** que garantice el cumplimiento de requisitos existentes en relación con los procesos y con los productos estadísticos, y la calidad de los sistemas estadísticos en su conjunto. Un marco de garantía de la calidad permite a las organizaciones nacionales e internacionales evaluar, comparar y mejorar las estadísticas de forma sistemática.

Un marco de garantía de la calidad estadístico verifica las siguientes características:

- Proporciona un mecanismo sistemático para la identificación y resolución de problemas de calidad de forma continua, y para maximizar la interacción entre personal estadístico;
- Ha sido aceptado como una parte esencial de la infraestructura de la oficina estadística;
- Es la base para la creación y el mantenimiento de una cultura de calidad dentro de la oficina y es una fuente valiosa de material de referencia para la formación;
- Hace transparente el proceso por el cual se asegura la calidad y refuerza la imagen de la oficina como un proveedor fiable de estadísticas de buena calidad;
- Facilita el intercambio de ideas sobre la gestión de la calidad con otros productores de estadísticas nacionales e internacionales.

Hoy en día la importancia del marco de garantía de la calidad se refleja en que la mayoría de los organismos internacionales han desarrollado su propio marco, lo mismo que la mayoría de los institutos de estadística de los países desarrollados. Algunos ejemplos son:

---

- La división de Estadística de Naciones Unidas dispone de dos marco de garantía de la calidad:
  - *National Quality Assurance Framework* (NQAF). Adoptado en marzo de 2019 sustituye la versión anterior que databa de 2012.  
<https://unstats.un.org/unsd/methodology/dataquality/>
  - *United Nations Statistics Quality Assurance Framework* (UNSQAF). Para ayudar a las agencias de Naciones Unidas (FAO, ITU, UNCTAD y UNIDO) a desarrollar su propio marco de garantía de la calidad.  
<https://unstats.un.org/unsd/unsystem/Documents-March2017/UNSystem-2017-3-QAF.pdf>

- La Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha desarrollado un marco que se centra en mejorar la calidad de los datos recogidos, compilados y difundidos por la organización mediante la mejora de los procesos y la gestión de la misma.

<https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm>

- Statistics Canada.  
<https://www150.statcan.gc.ca/n1/en/catalogue/12-586-X>
- Australian Bureau of Statistics.  
<https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+data+quality+framework>

De igual forma, el Sistema Estadístico Europeo también ha desarrollado su propio marco. El CBPEE es uno de los pilares del Marco Común de Calidad de las Estadísticas Europeas, junto con:

- El marco de garantía de calidad del Sistema Estadístico Europeo. Como apoyo y complemento al Código de Prácticas existe el denominado ESS QAF que proporciona un conjunto de métodos, herramientas y actividades para la implantación del Código.
- Los principios generales de gestión de la calidad (como la interacción permanente con los usuarios, el compromiso de liderazgo, los acuerdo de cooperación, la satisfacción del personal, la mejora continua, la integración y la armonización).

El marco de garantía de la calidad del Sistema Estadístico Europeo (*Quality Assurance Framework of the European Statistical System*, ESS QAF) (Eurostat 2019) es un documento de apoyo cuyo fin es ayudar en la implementación del CBPEE.

El ESS QAF presenta un conjunto de métodos, herramientas y buenas prácticas que, o bien ya están siendo usadas, o que se sugiere incorporar en el Sistema Estadístico Europeo. Su objetivo es acompañar al CBPEE proporcionando orientación y ejemplos en forma de métodos y herramientas más detalladas, así como de buenas prácticas para los principios e indicadores incluidos en el CBPEE. Algunos indicadores del CBPEE son



ellos mismos recomendaciones, por lo que las actividades de apoyo, los métodos y las herramientas pueden ser más detallados con el fin de facilitar la implementación del indicador.

Estos métodos, herramientas y buenas prácticas sugeridas se identifican a nivel institucional y a nivel de proceso/resultado, reflejando el nivel más adecuado para su adopción, aplicación y/o uso. En el documento, evolucionan desde una perspectiva general a una descripción más concreta y detallada. Es de resaltar que un método, herramienta o buena práctica específica (con alguna ligera modificación) puede ser apropiada para varios indicadores, y que se hace a propósito el hecho de señalarlo de forma repetitiva.

Las actividades, métodos y herramientas incluidas en el ESS QAF están diseñadas de forma que no dependan de soluciones organizacionales existentes en los países miembros, sino que sean genéricas, y a menudo van acompañados de ejemplos específicos que han funcionado bien en algunos países.

De acuerdo con la naturaleza autorreguladora del ESS QAF, estos métodos, herramientas y buenas prácticas no son de naturaleza obligatoria y por tanto, no son vinculantes para las autoridades estadísticas. Los métodos, herramientas y buenas prácticas incluidas demuestran cómo el CBPEE puede implementarse en el trabajo diario de un estadístico, teniendo en cuenta las circunstancias nacionales.

La primera versión del ESS QAF se publicó en 2012 e incluía únicamente los principios cuatro y del siete al quince. En 2015 se publicó una nueva versión obtenida a partir de las conclusiones del segundo ejercicio del *peer review*, véase la Sección 18.5. En esta nueva versión se desarrollaron los métodos y procesos para evaluar los principios cinco y seis y se propuso un ajuste de algunas actividades, métodos y herramientas para otros principios.

La última versión del ESS QAF es la 2.0 y se aprobó en 2019 con el fin de que se adaptara a la versión del CBPEE de 2017. El ESS QAF permanece abierto a futuras actualizaciones y seguirá de cerca la evolución del CBPEE en el futuro.

Para cada indicador de cada uno de los dieciséis principios del CBPEE pueden indicarse únicamente métodos a nivel institucional y/o métodos a nivel de proceso/resultado. Por ejemplo:

- Indicador 1.1: La independencia de los Institutos Nacionales de Estadística y de Eurostat frente a las injerencias externas, tanto políticas como de otro tipo, en el desarrollo, la elaboración y la difusión de estadísticas está establecida por ley y garantizada para otras autoridades estadísticas.

Métodos institucionales

### **1. Independencia profesional de los Institutos Nacionales de Estadística**

---

**y de Eurostat.** La ley estadística nacional incluye una disposición sobre el principio de independencia profesional de los Institutos Nacionales de Estadística. El Reglamento sobre Estadísticas Europeas incluye una disposición sobre el principio de independencia profesional de los Institutos Nacionales de Estadística y de Eurostat.

2. **Independencia profesional de otras autoridades estadísticas.** La legislación garantiza la independencia profesional de otras autoridades estadísticas.
  3. **Elaboración de legislación.** Se aprovechan directrices internacionales y/o buenas prácticas de otros países cuando se elaboran leyes relacionadas con las estadísticas.
  4. **Medidas legislativas.** Para promover/fomentar la legislación relacionada con las estadísticas, se dispone de asesoramiento y de unas estructuras administrativas adecuadas en el procedimiento legislativo del país para ayudar a la aprobación de estas medidas.
  5. **Código de ética profesional.** Existe un código de ética profesional y es público.
  6. **Sensibilización del personal.** Todo el personal estadístico conoce el código de ética profesional para estadísticas oficiales.
- Indicador 12.3: Las revisiones son analizadas de forma regular con el fin de mejorar las fuentes de datos, los procesos estadísticos y los resultados.

#### Métodos institucionales

1. **Procesos sobre el análisis de revisiones.** Se dispone de procesos sobre el análisis de los efectos de las revisiones en la precisión y la relevancia de los resultados. Los resultados de los análisis se utilizan para mejorar la calidad de las fuentes de datos, los procesos estadísticos y los resultados.

#### Métodos de tipo proceso/resultado

1. **Análisis de las revisiones.** Se analizan las revisiones. Esto sirve para mejorar las fuentes de datos, los procesos estadísticos y los resultados. El resultado de los análisis se usa para ajustar el ciclo de producción.
  2. **Indicadores de calidad de revisiones.** Indicadores de calidad de las revisiones (por ejemplo, el tamaño y la dirección de las revisiones y sus medias) se calculan de forma periódica de acuerdo a los estándares vigentes y son públicos.
- Indicador 14.4: Se realizan comparaciones y contrastes de estadísticas de distintas fuentes de datos y con distintas periodicidades.

#### Métodos de tipo proceso/resultado

1. **Comparaciones de los resultados estadísticos con datos relacionados.** Se realizan comparaciones de resultados estadísticos con otras estadísticas
-

o con datos administrativos que proporcionan información similar sobre el mismo dominio/fenómeno.

2. **Identificación y explicación de divergencias.** Se identifican las divergencias en los resultados estadísticos de distintas fuentes de datos y las razones son explicadas públicamente y de forma clara.
3. **Reconciliación de resultados estadísticos.** Los resultados estadísticos son contrastados siempre que sea posible.

## 18.4 La calidad en los productos y en los procesos estadísticos

La información estadística es esencial para que una sociedad organizada funcione. La falta de calidad afectaría seriamente a los procesos de toma de decisiones, la distribución de recursos y la habilidad de gobierno, empresas, instituciones, y el público en general de entender la realidad social y económica del país.

La credibilidad de una organización que produce estadísticas oficiales depende de varios factores, estando la calidad de los datos entre ellos. Las dimensiones de la calidad son de suma importancia para un instituto de estadística. Si una organización no es capaz de producir datos de calidad, tanto los usuarios como los informantes perderán pronto su confianza en ella, haciendo que su misión sea más difícil de conseguir. En consecuencia, un conjunto de mecanismos rigurosos que gestionen los asuntos relacionados con la calidad es vital en un INE.

En el caso de productos estadísticos producidos con cierta regularidad (como son las operaciones estadísticas mensuales o trimestrales), la evaluación debería de tener lugar, en teoría, en cada iteración, de forma que se pueda determinar si es necesarios hacer cambios en futuras iteraciones. Sin embargo, en el caso de procesos estadísticos regulares y consolidados, se considera que es suficiente con realizar esta evaluación una vez al año.

La consecuencia de esta evaluación debe de ser la identificación de procesos a mejorar así como la elaboración de un plan de mejora que incluya las acciones a realizar así como un calendario.

Cuando se piensa en la calidad de la estadística, en muchas ocasiones en lo que realmente se piensa es en la calidad de los productos estadísticos. Esta calidad se puede medir de varias formas: no sólo de forma cuantitativa sino también de forma cualitativa. Si bien lo más común es pensar en una serie de indicadores numéricos que proporcionen información sobre las distintas dimensiones que afectan a los productos. Debido a la importancia que tiene para Eurostat la calidad de las estadísticas, existen una guía<sup>3</sup> con conceptos, definiciones y ejemplos con el fin de ayudar en la medición de indicadores

---

<sup>3</sup><https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf>

comparables.

En la siguiente tabla se muestran los principales indicadores relacionados con las distintas dimensiones de la calidad en los productos.

Dimensión	Breve descripción
Relevancia	Tasa de completitud de los datos: número de datos proporcionados en función de los requeridos por el Plan Estadístico Nacional/Reglamento Europeo.
Precisión	<ul style="list-style-type: none"> <li>- Error de muestreo: que se puede proporcionar en términos relativos <math>CV(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}</math> o en forma de intervalo de confianza <math>[\hat{\theta} \pm t \cdot \sqrt{\hat{V}(\hat{\theta})}]</math></li> <li>- Tasa de sobrecobertura: proporción de unidades accesibles en el marco que no pertenecen a la población objetivo.</li> <li>- Tasa de falta de respuesta: que puede ser total, <math>NRr_w = 1 - \frac{\sum_R w_j}{\sum_R w_j + \sum_{NR} w_j + \alpha \sum_Q w_j}</math> y parcial.</li> <li>- Revisión de los datos: MAR, RMAR y MR. Si denotamos por <math>n</math> la longitud de datos a revisar, <math>X_{Lt}</math> el último valor y <math>X_{Pt}</math> el valor previo, las fórmulas son  <math display="block">MAR = \frac{1}{n} \sum_{t=1}^n  X_{Lt} - X_{Pt} </math> <math display="block">RMAR = \frac{\sum_{t=1}^n  X_{Lt} - X_{Pt} }{\sum_{t=1}^n  X_{Lt} }</math> <math display="block">MR = \frac{1}{n} \sum_{t=1}^n (X_{Lt} - X_{Pt})</math></li> <li>- Tasa de imputación: unidades imputadas sobre el total de la muestra.</li> </ul>
Oportunidad y puntualidad	<ul style="list-style-type: none"> <li>- Demora de los primeros resultados: número de días/semanas/meses desde el último día del periodo de referencia y la publicación de los primeros resultados.</li> <li>- Demora de los resultados finales.</li> <li>- Puntualidad.</li> </ul>
Comparabilidad y coherencia	<ul style="list-style-type: none"> <li>- Coeficiente de asimetría para las estadísticas espejo.</li> <li>- Longitud de las series de datos comparables.</li> </ul>
Accesibilidad y claridad	<ul style="list-style-type: none"> <li>- Número de consultas a las tablas de datos.</li> <li>- Número de consultas a los metadatos.</li> <li>- Tasa de completitud de los metadatos.</li> </ul>

Toda la información sobre la calidad de los productos debería de estar disponible para

los usuarios a través de los informes de metadatos de referencia en la página web. Estos informes también se deberían de proporcionar a Eurostat. La estructura de estos informes sigue un estándar del SEE. Se basa, esencialmente, en los componentes de calidad de los productos y además incluyen información sobre los procesos que han dado lugar a dichos productos.

En cualquier proceso de producción, la calidad debe de ser tenida en cuenta en todas sus fases. Esto no es distinto en la estadística. Las organizaciones estadísticas deben modelizar su proceso estadístico de forma que no sólo funcione sin problemas, sino también de forma que integre y utilice las dimensiones de calidad en cada fase.

El GSBPM ([UNECE 2019](#)) es un modelo para los procesos de producción que las oficinas de estadística siguen para producir estadísticas oficiales, conceptualizado en términos de fases y subprocesos asociados. Con el fin de supervisar la calidad de los procesos de producción estadística de cada una de estas fases, la UNECE en colaboración con otros organismos (Eurostat, Statistics Canada, etc.) ha desarrollado una serie de indicadores para productores de datos, metodólogos y gestores de proyectos. Los indicadores de calidad en los procesos estadísticos ([UNECE 2018](#)) juegan un papel fundamental en la gestión de la calidad, permitiendo controlar y prevenir errores. Estos indicadores de calidad están asociados a los 44 subprocesos del GSBPM tanto para encuestas como para datos administrativos.

Los principios utilizados para desarrollar estos indicadores de los procesos estadísticos son:

- Usar indicadores genéricos (no fórmulas) para reflejar la naturaleza el GSBPM como modelo de referencia;
- Ser consistente con los marcos de garantía de calidad internacionales existentes al seleccionar los indicadores de calidad y determinar la dimensión de calidad con la que están relacionados;
- Usar indicadores cuantitativos cuando sea posible;
- Usar indicadores cualitativos en la forma de *sí/no* o *mucho/medio/bajo* cuando sea apropiado;
- Permitir un cierto grado de redundancia mencionando los mismos indicadores en distintos subprocesos.

Ejemplo de indicadores de calidad asociados al subproceso 4.2 Inicializar recogida/obtención

---

Dimensión de calidad	Indicador
Confidencialidad estadística y seguridad	Riesgo de incumplimiento de mientras los datos están siendo transferidos.
Suficiencia de recursos	Tasa de recursos humanos e informáticos necesarios alcanzada.
Solidez de la implementación	Tasa de éxito del personal de recogida para realizar las tareas después de la formación. Tasa de éxito de las pruebas de los sistemas de recogida tanto en condiciones normales como en situaciones extremas o de gran volumen de datos.
Oportunidad y puntualidad	Retraso entre el cierre esperado y real de los sistemas de recogida (incluyendo la transmisión de datos, sistemas de gestión de datos) y materiales de recogida (por ejemplo, cuestionarios, material de formación, etc.).

Además de los indicadores desarrollados por la UNECE, Statistics Canada también ha elaborado unas indicaciones que aseguren la calidad en cada una de las fases del GSBPM. Se dispone de indicaciones genéricas para todos los procesos estadísticos y también indicaciones específicas para procesos que utilicen registros administrativos, nuevas encuestas, integración de datos.

## 18.5 Sistemas de evaluación global de la calidad: auditorías, autoevaluación y revisiones por homólogos en las oficinas de Estadística

Se pueden distinguir dos niveles de evaluación:

- la evaluación como parte de la gestión global de la calidad;
- la evaluación de casos individuales de procesos estadísticos.

La evaluación como parte de la gestión global de la calidad está relacionada con distintos tipos de gestión de la calidad: la formación en calidad, la medición de la calidad, el sistema de auditorías, y la concienciación de la calidad. La gestión de la calidad conlleva la evaluación de procesos estadísticos para identificar posibles duplicidades o vacíos.

El marco de calidad en el SEE está constituido por:

- El Código de Buenas Prácticas de las Estadísticas Europeas;
- El marco de garantía de la calidad del Sistema Estadístico Europeo;
- Los principios generales de gestión de la calidad.

La evaluación global de la calidad puede ser llevada a cabo por auditorías, autoevaluación o revisiones por homólogos en las oficinas de Estadística. Actualmente en algunos

países tienen lugares auditorías externas que evalúan de forma global la calidad de sus procesos estadísticos. Por otro lado, en muchos INEs se realizan ejercicios de autoevaluación utilizando para ello una serie de indicadores de calidad y de rendimiento <sup>4</sup>.

El QAF identifica actividades, métodos y herramientas que permiten evaluar el cumplimiento de los países del CBPEE, de sus principios y sus indicadores. La implementación del CBPEE se lleva a cabo con un enfoque de autorregulación voluntario. La principal herramienta para evaluar el progreso de su implementación son los *ESS peer reviews*, o revisiones por homólogos en las oficinas de Estadística (Eurostat 2021).

Su objetivo es fomentar la integridad, la independencia y la responsabilidad de las autoridades estadísticas que constituyen el SEE. La primera ronda de *peer reviews* tuvo lugar en 2006-2008, la segunda en 2014-2015 y la tercera en 2021-2023. Cubren tanto a los países miembros de la UE como a los de la EFTA. Las principales características son:

- En la primera ronda se analizó el cumplimiento de los principios 1 a 6 y 15, y la coordinación de los INEs. A partir de la segunda se evalúan todos los principios.
- Organizados por Eurostat. Los equipos están formados por experto(s) de Eurostat y de los INEs. De cara a la tercer ronda, la ESGAB recomienda que algún miembro sea externo al SEE.
- Se basan en las respuestas a un cuestionario de autoevaluación y en las conclusiones de la visita de los expertos al país.
- En la visita, los expertos se reúnen con empleados de los INEs (personal de las últimas promociones, directivos y personal intermedio) y con usuarios (investigadores, periodistas, principales usuarios).
- A partir de la segunda ronda también se incluyeron Otras Autoridades Nacionales (OANs), que pueden rellenar un cuestionario más sencillo que el de los INEs. La selección de las OANs se lleva a cabo teniendo en cuenta su importancia en la producción de estadísticas europeas.
- En algunos países también los Bancos Centrales Nacionales participaron en el proceso.

El equipo de *peer reviews* elabora un informe después de la visita a cada país y cada país tiene que elaborar un listado con las acciones de mejora y un plan de implementación. Periódicamente se revisa el cumplimiento de este plan de implementación.

Después de cada ronda de *peer reviews* también tiene lugar una revisión del CBPEE con el fin de realizar una actualización del mismo usando para ello las principales conclusiones obtenidas durante la ronda.

---

<sup>4</sup>Traducimos *Quality and Performance Indicators* como indicadores de calidad y de rendimiento.

Los *peer reviews* asumen ciertas funciones auditoras (formalización, estructura y ámbito predefinidos y listas de control detalladas y normas de evaluación) en relación con la evaluación del marco institucional y las prácticas de difusión de los INEs frente a los principios e indicadores del CBP. Proporcionan una oportunidad única de identificar estándares y dificultades comunes o carencias, el intercambio de buenas prácticas y de conocimientos en el Sistema Estadístico Europeo.

Los *peer reviews* contribuyen a que el Sistema Estadístico Europeo mejore el cumplimiento del Código y la calidad de las estadísticas europeas.

## Bibliografía

- Eurostat (2017). *Código de Buenas Prácticas de las Estadísticas Europeas*. URL: [https://ine.es/ine/codigobp/codigo\\_2017.pdf](https://ine.es/ine/codigobp/codigo_2017.pdf).
- (2019). *El marco de garantía de la calidad del sistema estadístico europeo*. URL: <https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>.
  - (2021). *Peer Review*. URL: <https://ec.europa.eu/eurostat/web/quality/peer-reviews>.
- UNECE (2018). *Indicadores de Calidad del GSBPM*. URL: <https://statswiki.unece.org/display/GSBPM/Quality+Indicators>.
- (2019). *The Generic Statistical Business Process Model v5.1*. URL: <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>.
-