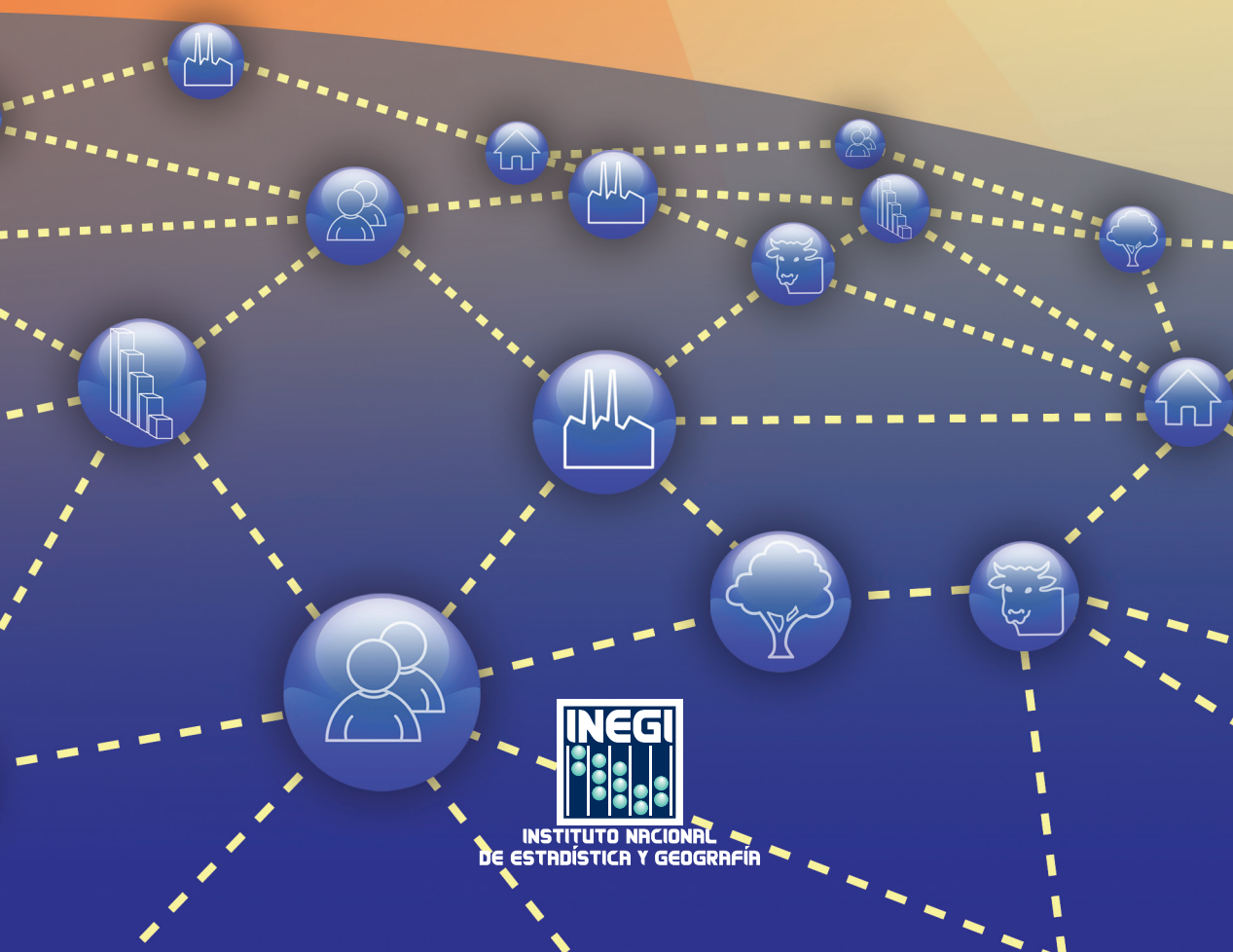
Instituto Nacional de Estadística y Geografía

Estadísticas basadas en registros

Aprovechamiento estadístico de datos administrativos

Anders Wallgren
Britt Wallgren



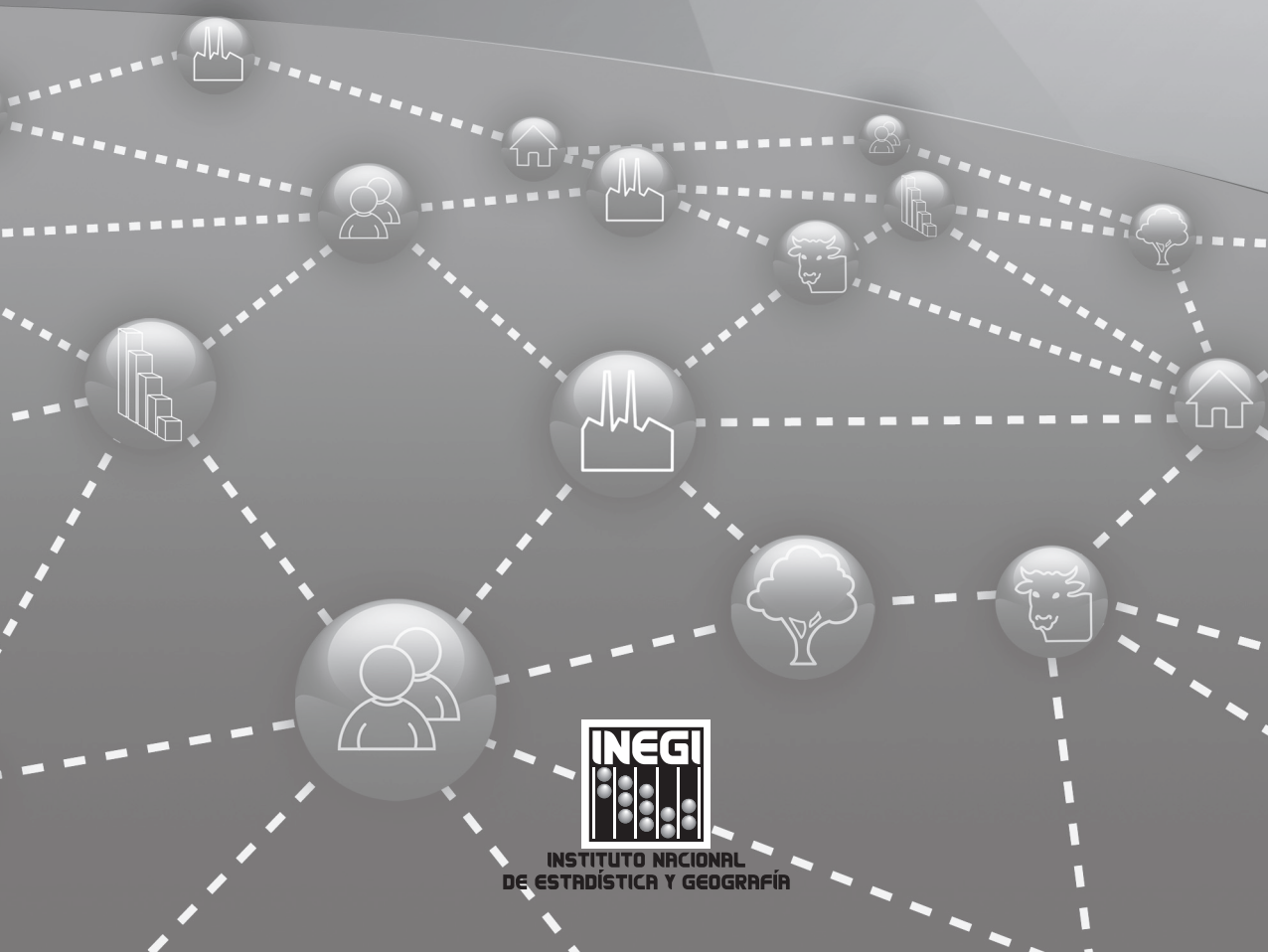
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Instituto Nacional de Estadística y Geografía

Estadísticas basadas en registros

Aprovechamiento estadístico de datos administrativos

Anders Wallgren
Britt Wallgren



INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Catalogación en la fuente INEGI:

310.01 Wallgren, Anders.
Estadísticas basadas en registros : aprovechamiento estadístico de datos administrativos / Anders Wallgren, Britt Wallgren.-- México : INEGI, c2016.

vii, 261 p.

1. Estadísticas - Registros. 2. Administración pública - Estadísticas - Metodología. I. Wallgren, Britt.

Conociendo México

01 800 111 4634
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa **@INEGI_INFORMA**

Agradecimientos

El aprovechamiento estadístico de los registros administrativos se ha considerado, desde hace varios años, como el mejor instrumento para reducir los costos en la generación de información y disminuir la carga para los informantes, utilizando datos que ellos entregan a diversas instancias del gobierno.

En este sentido, diferentes organismos internacionales han realizado esfuerzos para generar capacidades en las oficinas de estadística, con el objeto de incrementar el uso de las fuentes administrativas y asegurar la calidad de los datos generados.

En el ámbito internacional el avance de los países nórdicos en la materia goza de gran reconocimiento, por ello la experiencia de la Oficina de Estadística de Suecia resulta de gran valor para América Latina, ya que aporta elementos que permitirán mejorar las capacidades de los generadores de datos estadísticos de la región.

Anders Wallgren y Britt Wallgren en su empeño por contribuir al desarrollo de las estadísticas basadas en registros han compartido su experiencia a través de este libro y han puesto todo su esfuerzo y dedicación para que el conocimiento se transmita de la mejor manera a todos los países de América Latina, por lo cual les agradecemos ampliamente.

El INEGI también agradece a la editorial Wiley & Sons Ltd por la confianza para permitirnos hacer la traducción de un documento tan valioso. Dicha traducción no sería posible sin el apoyo del Banco Interamericano de Desarrollo, de manera particular de Janine T. Perfit quien ha respaldado el proyecto. La traducción ha estado a cargo de Virginia Aguirre Muñoz, a quien reconocemos su capacidad y entrega.

El producto de dichos esfuerzos es la presente obra que estamos seguros será de utilidad a todos los países de América Latina.

Contenido

Prefacio

Las encuestas basadas en registros se han vuelto cada vez más comunes no sólo en las oficinas nacionales de estadística, sino también en empresas y otras organizaciones que usan datos administrativos propios para generar estadísticas sobre producción, ventas y salarios, por ejemplo.

Aunque las estadísticas basadas en registros constituyen la forma más usual de estadística, hasta ahora no existe una teoría bien establecida en el campo. La ausencia de términos o principios de uso común ha dificultado aún más el desarrollo tanto de esa modalidad de estadística como de la metodología estadística para los registros. Como consecuencia de lo anterior, se aplican métodos ajustados a cada caso en vez de una teoría que goce de aceptación general.

En muchos países se investiga la posibilidad de un mayor aprovechamiento de los datos administrativos para propósitos estadísticos. Por un lado, es necesario reducir la carga y los costos de la obtención de respuestas; por el otro, el aumento de la no respuesta en los censos y las encuestas por muestreo también exige esta nueva estrategia, para lo cual se requiere establecer métodos estadísticos adecuados para las encuestas basadas en registros.

Desde hace más de diez años, hemos dedicado gran parte de nuestro trabajo en el Departamento de Investigación y Desarrollo de la Oficina de Estadística de Suecia al estudio de las estadísticas basadas en registros y sus requerimientos mediante el análisis del sistema de registros estadísticos de esa dependencia. Asimismo, hemos colaborado con varias empresas manufactureras y examinado sus datos administrativos en lo concerniente a su gestión, experiencias que aprovechamos para este libro.

Uno de nuestros propósitos consiste en describir el sistema de registros y analizar cómo debe funcionar, presentando las posibilidades que ofrece un sistema funcional. Otro propósito es estructurar y describir una metodología estadística para el aprovechamiento de registros que siente las bases para crear el sistema. En varias ocasiones sugerimos una nueva terminología. Este libro puede considerarse un primer paso hacia una forma más sistemática de trabajar con cuestiones estadísticas relacionadas con registros. Esperamos que sirva para impulsar el desarrollo de esta importante área y aliente la revisión y el mejoramiento de los métodos de trabajo actuales. Cabe señalar que este desarrollo necesario, por ejemplo para asegurar una mayor consistencia entre las diferentes encuestas basadas en registros, podría representar una tarea de muchos años.

El intercambio de puntos de vista con el profesor Carl-Erik Särndal fue fundamental para la elaboración de este libro. Su vasta experiencia en oficinas de estadística de varios países y sus conocimientos como especialista en encuestas por muestreo fueron de enorme utilidad. Además, alrededor de 50 funcionarios de la Oficina de Estadística de Suecia leyeron y comentaron diferentes partes de la primera versión, en sueco, de esta obra. Por otro lado, entrevistamos a varias personas en busca de material para los distintos ejemplos y las secciones metodológicas.

Este libro se publicó originalmente en sueco en 2004. En la Oficina de Estadística de Suecia, se ha usado en varios círculos de estudio que, además de estimularnos, nos ayudaron a preparar la versión en inglés, en la que se basa esta edición en castellano.

Al preparar esta versión tuvimos en cuenta los valiosos comentarios y sugerencias de cinco revisores anónimos contratados por John Wiley & Sons, Ltd.

Lars-Göran Lundell, de la Oficina de Estadística de Suecia, escribió la sección 11.6, *Sistemas informáticos para estadísticas basadas en registros*; con él examinamos varias partes a fin de incorporar al tema la perspectiva informática.

En suma, los comentarios y el apoyo de todas las personas mencionadas fueron muy valiosos en la elaboración de esta obra. Esperamos que las propuestas aquí planteadas estimulen el debate sobre el aprovechamiento de datos administrativos para propósitos estadísticos.

Örebro, Suecia
Enero de 2007

Anders Wallgren
Britt Wallgren

Prefacio a la edición en español

Los países del norte de Europa empezaron a usar registros administrativos para generar estadísticas en los años sesenta. La legislación en materia estadística establecía condiciones favorables para ese aprovechamiento, lo que explica por qué hoy las oficinas de estadística nórdicas pueden utilizar grandes cantidades de datos administrativos para fines estadísticos.

En muchos países, de manera particular en América Latina, la legislación en materia estadística no establece condiciones favorables para el aprovechamiento estadístico de datos administrativos, por lo que cambiar el sistema de producción de estadísticas a un sistema basado en registros llevará décadas. Mientras tanto, el trabajo de las oficinas de estadística será mejorar de manera gradual los sistemas administrativos, para que en el futuro los países puedan usar datos administrativos en un esquema estandarizado a nivel internacional. Por lo tanto, contar desde un principio con una comprensión clara de la experiencia nórdica facilitará y acelerará el proceso en los países que promuevan un mejor aprovechamiento de los registros administrativos con propósitos estadísticos.

Este libro es el resultado de la colaboración establecida en 2011 por iniciativa de José Antonio Mejía Guerra del Instituto Nacional de Estadística y Geografía de México y ha recibido el apoyo del Banco Interamericano de Desarrollo. Virginia Aguirre y Eric Manuel Rodríguez Herrera se ocuparon de la difícil e importante labor de la traducción del inglés al español. Estamos muy agradecidos por el trabajo que realizaron.

Örebro, Suecia
Abril de 2012

Anders Wallgren
Britt Wallgren

CAPÍTULO 1

Encuestas basadas en registros: introducción

En este capítulo y el siguiente presentaremos una serie de conceptos y principios que usaremos a lo largo del libro en relación con las *encuestas basadas en registros*, es decir, encuestas que usan como fuente de información datos de registros administrativos. Estos conceptos y principios constituyen el fundamento para elaborar una teoría sobre esta modalidad de encuesta.

Las encuestas basadas en registros son comunes dentro de empresas y otras organizaciones, que usan los datos de sus propios sistemas administrativos para generar estadísticas sobre producción y ventas, por ejemplo. Este tipo de encuestas también es común en las oficinas nacionales de estadística de los países escandinavos, donde se aprovechan muchos registros administrativos para producir estadísticas oficiales.

Analizaremos principalmente las encuestas basadas en registros que realizan las oficinas nacionales de estadística. Hay un interés creciente en este campo porque en muchos países se recurre cada vez más a los datos administrativos para fines estadísticos y, por ende, hay una mayor demanda de una teoría sobre este tipo de encuestas.

Nuestro objetivo es presentar métodos y principios estadísticos de interés general, pero nos valdremos de experiencias escandinavas y estudios de caso de la Oficina de Estadística de Suecia para ilustrar estos aspectos metodológicos generales.

1.1 ¿ES NECESARIA UNA TEORÍA SOBRE LAS ENCUESTAS BASADAS EN REGISTROS?

Las oficinas nacionales de estadística publican tres tipos de estadísticas: basadas en encuestas por muestreo, basadas en censos y basadas en registros administrativos. Por lo general, sólo se distingue entre encuestas y censos, modalidades en las que predomina el trabajo de captación de datos, que es responsabilidad de la oficina de estadística.

En este libro abordaremos el tercer tipo de estadísticas, las basadas en registros, en las que en vez de captar datos mediante encuestas y censos, se adaptan y procesan registros administrativos de diferentes fuentes con el objeto de adecuarlos para propósitos estadísticos. Este tipo de encuesta recibe el nombre de *encuesta basada en registros.*

Las encuestas por muestreo se apoyan en métodos derivados de la *teoría del muestreo.* Esta teoría, firmemente establecida, se formuló tanto en los círculos académicos como en las oficinas de estadística. Consta de términos y principios tan conocidos que permiten a quienes trabajan con encuestas por muestreo comunicarse e intercambiar experiencias sin

dificultad. Asimismo, en la bibliografía y las publicaciones especializadas se plantean y difunden metodologías para el muestreo y la estimación.

Los censos con una captación de datos propia tienen como fundamento una larga tradición de censos poblacionales y los datos recopilados por autoridades locales, escuelas y diferentes tipos de empresas. Los errores de medición, el diseño de los cuestionarios y la no respuesta son problemas metodológicos que también se presentan en las encuestas por muestreo. Los censos y las encuestas por muestreo guardan una estrecha relación en cuanto a metodología, incluso a menudo se considera que los censos son casos especiales donde la muestra es toda la población.

En lo sucesivo, llamaremos *estadísticas basadas en registros* a aquellas que se generan a partir de registros administrativos. Si bien es la forma más antigua y común de estadística, no existe una teoría sólida en el campo. No hay términos o principios bien establecidos, lo que dificulta aún más tanto el desarrollo de este tipo de estadística como la formulación de una metodología estadística para el aprovechamiento de los registros. Por consiguiente, *se usan métodos adaptados a cada caso en vez de una teoría generalmente aceptada*.

Un factor importante que explica esta deficiencia es que la estadística académica no incluye el campo de las encuestas basadas en registros. Dentro de las ciencias estadísticas, se entiende que la teoría estadística se compone de la *teoría de las probabilidades* y la *inferencia estadística*. La teoría del muestreo forma parte de esta escuela, no así las encuestas basadas en registros a partir de una enumeración total.

Por desgracia, hasta ahora las ciencias estadísticas no cuentan con ninguna teoría sobre sistemas estadísticos. Ni las oficinas de estadística, ni las empresas y organizaciones grandes realizan encuestas independientes con mucha frecuencia. Es más común la construcción de sistemas de información, lo que constantemente produce nuevos datos. Es necesaria una teoría de la estadística con el propósito de describir los principios generales y elaborar el aparato conceptual para dichos sistemas estadísticos. Esa teoría debe considerar las encuestas basadas en registros.

A continuación proponemos cuatro principios para el aprovechamiento adecuado de los registros administrativos:

Gráfico 1.1 Cuatro principios sobre el uso de datos administrativos

1. Una oficina de estadística debe tener acceso a los registros administrativos que mantengan las autoridades. Este derecho, al igual que la protección de la privacidad, debe estar amparado por ley.
2. Dichos registros administrativos se deben transformar en registros estadísticos. Durante este proceso, se usarán y compararán fuentes muy diversas.
3. Todos los registros administrativos se deben incluir en un sistema de registros coordinado. Este sistema asegurará que todos los datos se puedan integrar y usar eficazmente.
4. La consistencia en las poblaciones y las variables es necesaria para que las estimaciones de diferentes encuestas basadas en registros sean coherentes.

Usaremos estos principios a lo largo del libro e introduciremos paulatinamente la terminología de estadísticas basadas en registros que es necesaria para los debates.

En 1995, Statistics Denmark (la Oficina de Estadística de Dinamarca) publicó *Statistics on Persons in Denmark: A Register-based Statistical System*, donde se hace un análisis

sistemático del trabajo estadístico con registros y se explica cómo diseñar un sistema de registros bien estructurado.

En el presente libro, nos basamos en esa publicación danesa y la complementamos. En los siguientes capítulos introduciremos varios conceptos y principios de la estadística basada en el aprovechamiento de registros. En el *Glosario* se compilan todos estos conceptos y términos. Nuestro objetivo es que todas las personas que trabajan en la elaboración de encuestas basadas en registros compartan esa terminología

1.2 ¿QUÉ ES UNA ENCUESTA ESTADÍSTICA?

El punto de partida para cualquier encuesta es una serie de preguntas relacionadas con un área de interés específica. El propósito de una encuesta es tratar de dar respuesta a esas preguntas. El proceso de una encuesta puede describirse en mayor o menor detalle. De manera esquemática, el trabajo consta de las siguientes fases:

1. Determinación de los objetivos de investigación y planeación de la encuesta.
2. Captación y procesamiento de los datos.
3. Estimación, análisis de datos y presentación de resultados.

En una oficina nacional de estadística, es común trabajar con encuestas, que se repiten cada año, trimestre o mes. Con estas encuestas, el trabajo se concentra en las fases 2 y 3. Sin embargo, estas encuestas también han pasado por la fase de determinación de objetivos y planeación, aunque fuera mucho tiempo antes.

Una encuesta independiente puede ser un proyecto en que la oficina de estadística debe llevar a cabo toda la encuesta, lo cual supone trabajar en las tres fases. Sin embargo, en muchos de estos proyectos, el cliente se ocupa de las fases 1 y 3, y la oficina de estadística sólo de la 2.

La fase 2 de una encuesta, la captación de datos, se puede llevar a cabo de diferentes maneras:

a. Mediante una captación propia usando una *encuesta por muestreo.*
   Ejemplo: En muchos países se realiza una encuesta sobre fuerza laboral. Cada mes se toma una muestra nueva, con una nueva captación de datos y nuevos informes.
b. Mediante una captación propia usando un censo.
   Ejemplo: El censo tradicional de población y vivienda, en el que todos los hogares y los propietarios de vivienda son entrevistados o se les pide que respondan un cuestionario, que después procesa la oficina nacional de estadística.
   Como los censos tienen como resultado la creación de un registro, los microdatos de los censos también se incluyen en el sistema de registros estadísticos y, por tanto, pueden constituir el punto de partida de las encuestas basadas en registros.
c. Mediante los microdatos existentes para realizar una encuesta basada en registros.
   Los microdatos se refieren a datos sobre *objetos* individuales. Los registros administrativos o estadísticos que contienen datos referidos, por ejemplo, a personas o empresas se usan para el propósito de realizar una encuesta basada en registros.
   Ejemplo: En la sección 1.4 presentaremos dos ejemplos de cómo se crean registros estadísticos para satisfacer las necesidades de las diferentes encuestas basadas en registros.

Como estos tres tipos de encuesta difieren en cuanto a metodología, conviene diferenciarlos conceptualmente. Las encuestas por muestreo, los censos y las encuestas basadas en registros son los tipos más importantes de encuestas en una oficina nacional de estadística.

Una población estadística consta de *N objetos* o *unidades* o *elementos*. De estos tres sinónimos, usaremos el término *objeto*.

1.3 ¿QUÉ ES UN REGISTRO?

Se mantiene un *registro administrativo* para almacenar el registro sobre *todos* los objetos por administrar. El proceso administrativo debe permitir la *identificación* de todos los objetos. La siguiente definición se aplica a los registros tanto administrativos como estadísticos:

> Un *registro* tiene la finalidad de ser una lista completa de los objetos que forman parte de un grupo de objetos o una población específicos. Sin embargo, pueden faltar datos sobre algunos registros a causa de deficiencias en la calidad. Los datos sobre la identidad de un objeto también deben estar disponibles, de modo que el registro se pueda actualizar y ampliar con nuevos valores de variable para cada objeto. Por ende, las características más importantes de un registro son el listado completo y las identidades conocidas.

Las identidades utilizadas en el procesamiento de registros pueden ser números de identificación únicos dentro de un sistema administrativo nacional o un número de identificación en un subsistema con claves para las identidades en otros sistemas. También es posible usar identidades definidas por el nombre, el domicilio, la fecha de nacimiento o el lugar de nacimiento, por ejemplo.

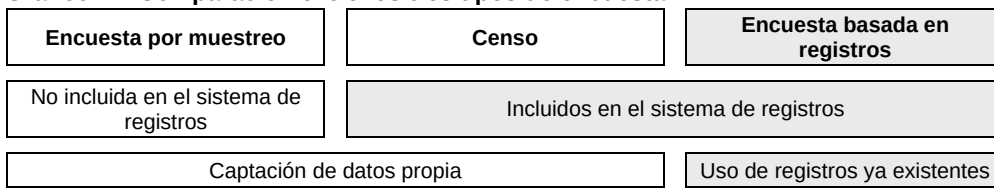
Estas identidades se usarán en un emparejamiento exacto de los objetos en diferentes registros, con objeto de hallar objetos idénticos o relacionados en dos registros.

Un *registro estadístico* se basa en datos de registros administrativos procesados para su adecuación a propósitos estadísticos. El procesamiento que transforma los datos administrativos en registros estadísticos plantea interesantes cuestiones metodológicas que veremos más adelante.

El término *registro estadístico* se aplica a los registros que forman parte de un sistema de registros estadísticos de una oficina de estadística u otra organización. Estos registros pueden estar basados en un censo realizado por una dependencia o en registros administrativos de autoridades u organizaciones ajenas a la oficina de estadística.

La captación de datos en una encuesta por muestreo no da origen a un registro, pues los microdatos sobre la muestra sólo corresponden a una pequeña parte de la población encuestada. En el gráfico 1.2 se comparan los tres tipos de encuesta predominantes en las oficinas nacionales de estadística.

Gráfico 1.2 Comparación entre los tres tipos de encuesta



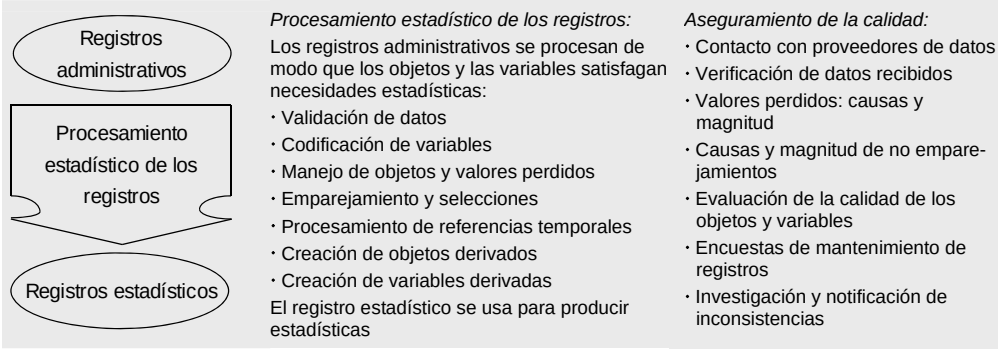
El término *estadística basada en registros* se refiere a las estadísticas que aprovechan encuestas basadas en registros. Cuando analizamos el sistema de registros, como en el capítulo 2, no distinguimos entre censos y encuestas basadas en registros. Sin embargo, cuando hablamos de problemas metodológicos, el término se refiere exclusivamente a encuestas basadas en registros.

1.4 ¿QUÉ ES UNA ENCUESTA BASADA EN REGISTROS?

Los registros administrativos se crean y entregan a una oficina nacional de estadística
La formación original de los datos corre a cargo de las autoridades y organizaciones, que adaptan las definiciones de los objetos y variables para propósitos administrativos. Cada autoridad efectúa controles, correcciones y otros procesos adecuados para sus finalidades administrativas. Los datos entregados por una autoridad a una oficina nacional de estadística pueden pasar por selecciones y procesos adicionales para satisfacer las necesidades de esa oficina. Las autoridades respectivas también tienen metadatos, es decir, información sobre las definiciones, la formación de datos y la calidad. Este tipo de información también es importante para quienes reciben los datos en la oficina de estadística.

¿Qué ocurre cuando los datos se entregan a una oficina de estadística como la Oficina de Estadística de Suecia?
En términos generales, no es conveniente generar estadísticas a partir de los registros administrativos recibidos originalmente porque éstos no se adaptan a los requisitos estadísticos. Es necesario validar los conjuntos de objetos, las definiciones de objetos y las variables, y muchas veces incluso llevar a cabo algún tipo de procesamiento a fin de que el registro satisfaga los requisitos estadísticos para objetos y variables. El procesamiento estadístico de los registros, que tiene como finalidad transformar uno o varios registros administrativos en un solo registro estadístico, debe basarse en una *metodología estadística para el aprovechamiento de registros* generalmente aceptada. Analizaremos con mayor detalle estos problemas metodológicos en los próximos capítulos. En el gráfico siguiente se muestran los diferentes elementos incluidos en el trabajo de metodología estadística.

Gráfico 1.3 De un registro administrativo a un registro estadístico



En las siguientes dos subsecciones, explicamos cómo se crean los registros estadísticos. Los ejemplos están tomados de la Oficina de Estadística de Suecia, pero sirven para ilustrar principios generales. Cada uno de estos registros se crea para satisfacer las necesidades de varias encuestas basadas en registros. Los ejemplos ilustran la manera en que se transforman los datos administrativos para satisfacer necesidades estadísticas y en que se aprovecha

el sistema de registros estadísticos de esta dependencia sueca cuando se crean registros estadísticos. La parte medular de una encuesta basada en registros es la creación de un registro apropiado.

1.4.1 Registro de Ingresos y Tributación de la Oficina de Estadística de Suecia

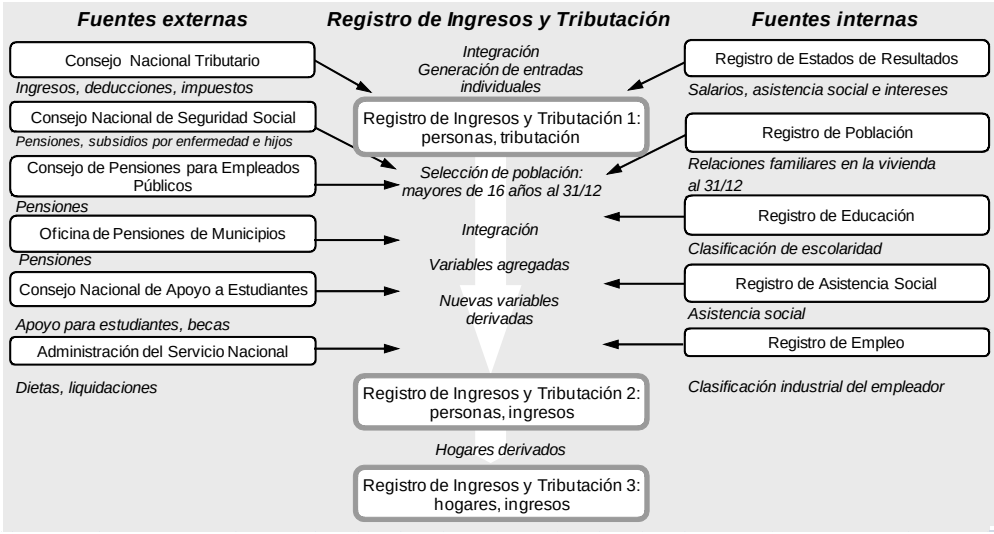
En este registro se usa una gran cantidad de fuentes administrativas y muchas variables administrativas para crear variables estadísticas importantes. Además de estas fuentes administrativas, es necesario recurrir al sistema de registros de la Oficina de Estadística de Suecia: el Registro de Población se utiliza para definir la población y se importan variables de clasificación importantes de otros registros del sistema.

1. *Formación de datos en la Consejo Nacional Tributario*
   La evaluación del ingreso anual se basa tanto en las declaraciones de impuestos de los contribuyentes como en las decisiones en materia de tributación de la autoridad hacendaria local. Los contribuyentes y la autoridad hacendaria usan los estados de resultados en lo relativo a salarios, subsidios por enfermedad e intereses de los que son responsables los empleadores, la oficina del seguro social y las compañías financieras. Por último, el Consejo Nacional Tributario (CNT) sueco compila esta información. Las declaraciones, los estados de resultados y las decisiones en materia tributaria pueden modificarse y complementarse. Por consiguiente, los datos de una sola persona pueden ser complejos.
2. *Entrega de microdatos al Registro de Ingresos y Tributación*
   Cada año, el CNT genera bases de datos que contienen información sobre la población del país. Los archivos de datos correspondientes a un año —que contienen alrededor de nueve millones de entradas, cada una con alrededor de 300 variables— se entregan directamente al Registro de Ingresos y Tributación en la Oficina de Estadística.
3. *Metadatos para el Registro de Ingresos y Tributación*
   Las entregas del CNT incluyen la descripción de las entradas con nombres y definiciones de las variables. Para interpretar los datos también se requieren los formularios de declaración de impuestos, los formularios de los estados de resultados, las decisiones en materia tributaria, las instrucciones para las declaraciones de impuestos y las instrucciones para los empleadores.
4. *Validación de datos*
   El Registro de Ingresos y Tributación recibe datos de once proveedores diferentes fuera y dentro de la Oficina de Estadística de Suecia. Los datos que vienen de fuera se deben validar y los generados en la Oficina de Estadística ya están validados. El contacto con los proveedores es importante para conocer los cambios en el sistema administrativo, lo que permite asegurar la calidad de las estadísticas a partir de registros —los cambios administrativos no se deben interpretar como cambios reales en los ingresos—.
5. *Emparejamiento y selecciones*
   Se debe procesar un gran número de registros para crear los diferentes subregistros incluidos en el Registro de Ingresos y Tributación. Las entradas de diferentes fuentes se emparejan usando números de identificación personal (NIP) y al mismo tiempo se lleva a cabo la agregación, es decir, se agregan todos los datos sobre estados de resultados correspondientes a una persona específica, de modo que se puedan integrar los ingresos que percibe esa persona derivados de su trabajo. Un tipo de procesamiento consiste en seleccionar personas de 16 años o mayores que también formaban parte de la población al 31 de diciembre.

6. *Creación de objetos derivados*
   Tener mayor información sobre ciertas relaciones ayuda a formar unidades familiares. Los adultos con relación de *adultos casados* o *que cohabitan y tienen hijos en común* son incluidos en la misma unidad familiar. Estas relaciones se indican mediante los números de identificación personal de los miembros de la familia, que son variables de referencia en los datos sobre tributación y en el Registro de Población de la Oficina de Estadística.
7. *Creación de variables derivadas*
   Se forma un gran número de variables derivadas en relación con el ingreso. Por ejemplo, se suman los montos salariales de los diferentes datos sobre percepciones para obtener el *ingreso por trabajo* de una persona. El ingreso total de cada persona por trabajo y capital más pagos de transferencias menos impuestos da el *ingreso disponible* de esa persona. En el caso de los hogares, se crean variables como *tipo de hogar*, *número de unidades de consumo* e *ingreso disponible*.

En el gráfico siguiente se muestra la forma en que el Registro de Ingresos y Tributación recibe los datos administrativos de diversas fuentes externas y algunos registros de la Oficina de Estadística de Suecia. El término *registro fuente* se refiere a las fuentes administrativas y los registros de esa dependencia que se usan para generar el nuevo registro. En la columna central se indican las distintas fases en que se usan los registros fuente para generar el nuevo registro estadístico.

Gráfico 1.4 Diferentes fuentes de datos para el Registro de Ingresos y Tributación



Este ejemplo nos muestra la importancia de los cuatro principios del gráfico 1.1. La Oficina de Estadística de Suecia tiene acceso a muchos registros administrativos con variables que describen los diferentes tipos de ingreso. El conjunto de objetos y las variables administrativas se procesaron para satisfacer las necesidades estadísticas. Se han usado muchas fuentes para generar un registro estadístico de ingresos con un vasto contenido. La población del registro de ingresos es consistente con otros registros estadísticos dentro del sistema.

El Registro de Ingresos y Tributación constituye una parte importante del sistema de registros de la Oficina de Estadística de Suecia. Sirve para describir la distribución del

ingreso y elaborar estadísticas regionales de los ingresos, además de constituir la base de los registros de ingresos longitudinales que usan los investigadores universitarios.

1.4.2 Registro longitudinal: escolaridad y mercado de trabajo

El Registro de Ingresos y Tributación se alimenta directamente de grandes cantidades de datos administrativos. Sin embargo, no sucede así con muchos registros importantes de la Oficina de Estadística de Suecia, pues se basan en registros estadísticos que ya existen en el sistema de registros. El ejemplo ilustra la manera en que pueden aprovecharse datos ya existentes de una manera novedosa y más avanzada luego de un procesamiento especial para adaptar los registros.

La incorporación de los jóvenes al mercado de trabajo tras finalizar sus estudios es hoy un área de gran interés para diferentes encuestas. Esas encuestas deben llevarse a cabo como *encuestas longitudinales*, donde se da seguimiento a grupos de personas durante un periodo de años. Si se llevan a cabo como encuestas por muestreo, cada año se toma una muestra entre las personas que concluyen un programa educativo específico y a cada muestra se le entrevista o se le solicita que responda un cuestionario anualmente durante un periodo de varios años, en este caso de siete años.

Este método de encuestar tiene sus desventajas, en parte porque la carga para los informantes es grande —las personas seleccionadas deben responder un gran número de preguntas cada año— y en parte porque la no respuesta aumentará paulatinamente con el paso del tiempo. Además, si no se cuenta con un registro adecuado de los programas educativos terminales, entonces es necesario seleccionar una muestra grande de personas de cierto grupo de edad para identificar a quienes han concluido el nivel medio superior o superior.

Un método de encuestar alternativo consiste en que la encuesta se apoye exclusivamente en registros ya existentes. La Oficina de Estadística de Suecia publica este tipo de estadísticas basadas en registros, que describen la transición al mercado de trabajo. Estas estadísticas utilizan fuentes administrativas; sin embargo, para generar estadísticas no basta con que las oficinas de estadística tengan acceso a fuentes administrativas. *También es necesario que tengan acceso a un sistema en operación de registros estadísticos.* El ejemplo que presentamos sirve para ilustrar características importantes de las estadísticas basadas en registros y de un sistema de registros.

En los gráficos siguientes se comparan personas con un nivel de escolaridad inferior y superior conforme tratan de ingresar al mercado de trabajo. Durante el periodo 1988–1993 se da seguimiento a seis cohortes integradas por todas las personas que concluyeron el nivel medio superior en 1987–1992 y se puede comparar su transición a un empleo remunerado con las seis cohortes correspondientes de egresados universitarios. Estos años fueron de particular interés por tratarse de un periodo en el que el mercado de trabajo pasó por un cambio radical. Se estudió a todas las personas pertenecientes a estas doce cohortes por medio de registros longitudinales, que posteriormente se analizaron.

Los círculos de los gráficos 1.5 y 1.6 representan el porcentaje de personas con empleo remunerado *un* año después de haber concluido sus estudios. Las curvas muestran la evolución del porcentaje de personas con un empleo remunerado dentro de cada cohorte.

Porcentaje de personas empleadas tras concluir sus estudios, 1987–1992

Gráfico 1.5 Nivel medio superior

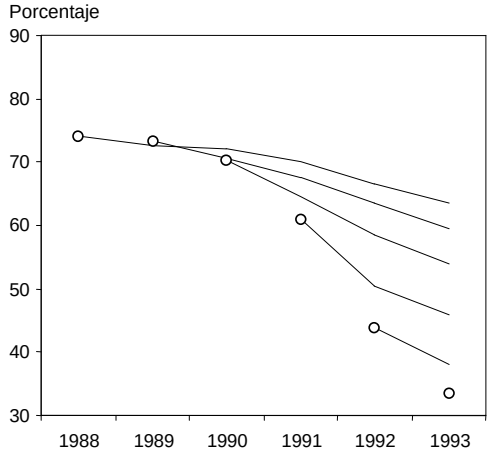
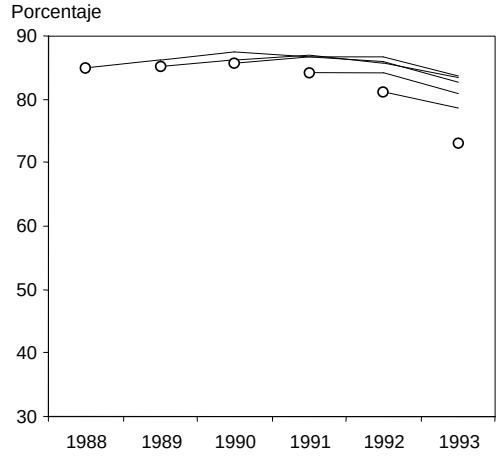


Gráfico 1.6 Egresados de la universidad



A principios de los años noventa, el mercado de trabajo sueco atravesó su crisis más grave desde los años treinta. Los gráficos 1.5 y 1.6 muestran que la contracción económica modificó las posibilidades de que los jóvenes se incorporaran al mercado de trabajo. El porcentaje con un empleo remunerado un año después de concluir la educación media superior cambió notoriamente en esos años. Para los egresados universitarios, las circunstancias no fueron tan desfavorables, lo que parece demostrar que los estudios superiores valieron la pena.

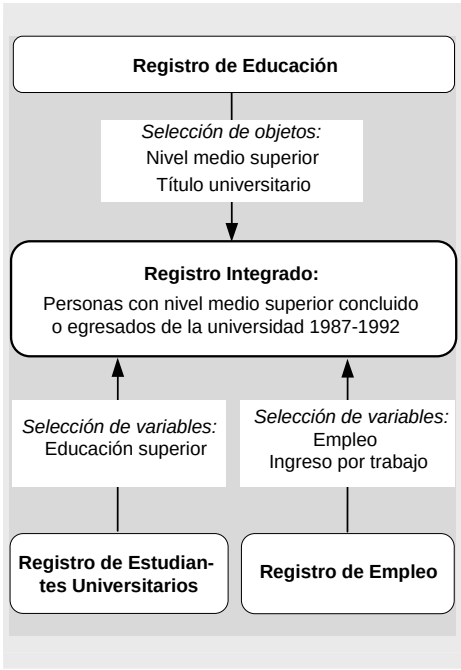
Este ejemplo se basa en datos longitudinales. Todos los años, cada persona se clasifica como trabajador en un empleo remunerado, estudiante de nivel superior o no perteneciente a ninguna de estas dos categorías. En el caso de los trabajadores en empleos remunerados, se registra tanto el ingreso anual por trabajo como el sector de actividad. También hay una clasificación por carrera o programa de estudios, sexo y región, lo que permite elaborar informes detallados. En este caso, el costo anual de este producto estadístico ascendió a 0.4 millones de coronas suecas (aproximadamente 50 000 dólares estadounidenses), la mitad de lo cual correspondió a la impresión.

¿Cómo se llevó a cabo esta encuesta basada en registros?

Los gráficos anteriores muestran la transición de los jóvenes de los estudios al mercado de trabajo. El registro longitudinal en el que se basan esos gráficos se creó de la siguiente manera:

- Combinando la información de tres registros de la Oficina de Estadística de Suecia para el periodo 1987–1993 (21 registros en total), se generó un nuevo *registro integrado*, enmarcado con un borde oscuro en el gráfico 1.7.
- Los objetos en el nuevo registro se crearon seleccionando ciertos objetos del Registro de Educación.
- Los valores de las variables se importan al nuevo registro *emparejando* los objetos del nuevo registro con los objetos correspondientes en el Registro de Educación, el Registro de Estudiantes Universitarios y el Registro de Empleo, como vemos en el gráfico siguiente.

Gráfico 1.7 Generación de un registro para la transición al mercado de trabajo



Registro de Educación
Basado en datos de escuelas o universidades.
Objeto: Persona.
Variables: Número de identificación personal, nivel de escolaridad, año de egreso.
Procesamiento: Selección de personas que concluyeron el nivel medio superior y egresados de la universidad.

Registro de estudiantes universitarios
Basado en datos de las universidades.
Objeto: Persona que estudia activamente.
Variables: Número de identificación personal, estudios superiores en curso.
Procesamiento: Emparejamiento con el grupo seleccionado. Las variables se seleccionan y transfieren al registro integrado.

Registro de empleo
Basado en estados de resultados, ingreso, población y registros empresariales.
Objeto: Persona.
Variables: Número de identificación personal, situación de empleo, ingreso por trabajo.
Procesamiento: Emparejamiento con el grupo seleccionado. Las variables se seleccionan y transfieren al registro integrado.

Ya hemos mencionado varios términos importantes. Al reunir o integrar información de varios registros, se genera un *registro integrado*. Debemos distinguir entre *selección de objetos* y *selección de variables*. *Emparejamiento* significa que se comparan los vínculos entre dos o más registros, lo que arroja una coincidencia o una no coincidencia.

Para generar el nuevo registro integrado, no se usaron nuevas fuentes administrativas, sino que se recurrió en mayor medida a los registros disponibles. El Registro de Empleo también se basa en los registros disponibles dentro del sistema de registros de la Oficina de Estadística de Suecia. Sin embargo, el Registro de Estudiantes Universitarios se conformó a partir de los datos administrativos que la Oficina de Estadística de Suecia recibe de otras autoridades.

Para crear y coordinar estos registros, se requiere un amplio procesamiento de los datos de varios registros. Esto significa que no basta con tener acceso a fuentes administrativas para producir estas estadísticas, lo cual sería imposible sin el sistema de registros estadísticos coordinado de la Oficina de Estadística de Suecia. Cuando se usa el sistema de registros, el emparejamiento exacto de las entradas de los diferentes registros se lleva a cabo con números de identificación estandarizados. En el ejemplo anterior, se usan los números de identificación personal (NIP).

1.5 SISTEMAS DE INFORMACIÓN ADMINISTRATIVA Y ESTADÍSTICA

El uso de datos administrativos para fines estadísticos no es exclusivo de las oficinas de estadística, ya sea de Suecia o de otros países. También es una práctica común en las grandes empresas y organizaciones. Por lo general, los sistemas administrativos se usan

como fuentes de información estadística y no hay gran diferencia entre el siguiente ejemplo sobre empresas y las estadísticas basadas en registros de una oficina nacional de estadística:

– Las estadísticas sobre personal y salarios dentro de una empresa se pueden producir mediante el sistema de gestión de personal.
– Las estadísticas sobre población e ingresos se producen en una oficina de estadística a partir de los datos del sistema de recaudación de impuestos de las autoridades hacendarias, para el registro de población y la determinación de la base impositiva.

Las encuestas basadas en registros son cada vez más comunes en las empresas y organizaciones. Tener conocimientos sobre sistemas de registro, estadísticas basadas en registros y calidad de los registros no es sólo necesario en una oficina nacional de estadística, sino también en un sentido más general, como vemos en los siguientes párrafos de una oferta de empleo:

Analista de mercados

Como analista del departamento de mercadotecnia, será un importante elemento para el crecimiento continuo de nuestra empresa. Se ocupará de gestionar y mejorar el uso de uno de los activos más valiosos de la empresa: nuestro registro de clientes.

Trabajará en análisis de campañas, redacción de informes, segmentación y aseguramiento de la calidad del registro. Mantendrá contacto con sistemas de registros externos y colaborará estrechamente con el gerente de mercadotecnia.

Algunos sistemas de información se establecen exclusivamente para fines estadísticos, como las encuestas sobre fuerza laboral, que se llevan a cabo en muchos países. Esos sistemas se pueden diseñar con estricto apego a principios estadísticos.

Otros sistemas de información se usan para fines tanto administrativos como estadísticos, lo que en ocasiones puede originar conflictos en relación con la estructura del sistema. En términos generales, estos sistemas están destinados sobre todo a fines administrativos, de modo que la información estadística es un subproducto.

Sin embargo, hay varias diferencias entre un sistema puramente administrativo y un sistema puramente estadístico. A continuación compararemos estos sistemas.

1.5.1 Propósitos diferentes

En un sistema administrativo, la información se usa como base cuando se toman medidas y decisiones administrativas que afectarán a los objetos del sistema.

Ejemplo: Se usa un sistema de gestión de personal para el pago de salarios.

En un sistema estadístico, la información sirve como base para el análisis, del cual se obtendrán conclusiones. Más adelante, estas conclusiones podrían ser el fundamento para decisiones en materia de políticas.

Ejemplo: Se usa un sistema salarial estadístico para estudiar la estructura salarial. ¿Cómo ha cambiado esto? ¿Cuáles son las diferencias en el salario mensual de las diferentes categorías de personal? Este análisis podría traer consigo un cambio en la política salarial, por ejemplo, un mejor pago para las mujeres.

1.5.2 Diferentes funciones para objetos individuales

En un sistema administrativo, se toman decisiones y medidas respecto a objetos individuales. Para ello, se recupera la información relacionada con ese objeto específico.

Ejemplo: Se paga un salario a todos los empleados de una empresa. Se verifica la información administrativa para poder calcular los salarios y los impuestos.

En un sistema estadístico, los objetos individuales no revisten interés por sí mismos. En un análisis estadístico las estimaciones agregadas se calculan y luego se comparan para grupos de objetos.

Ejemplo: Se calculan los totales de los salarios, los salarios promedio, la dispersión salarial, etcétera para las diferentes categorías de personal.

1.5.3 Criterios respecto a errores

Desde el punto de vista administrativo, ciertos elementos de la información sobre las personas deben ser absolutamente correctos, pero otros pueden ser aproximados. Puede haber errores desde el punto de vista estadístico, pero deben sujetarse a un control estricto a fin de reducir los errores, que pueden ser trascendentales para las conclusiones estadísticas. Se pueden aceptar errores para algunos datos, pero sólo si se considera que tendrán un efecto limitado.

Ejemplo: El número de identificación personal en un sistema de gestión de personal debe ser completamente correcto en relación con los salarios y la administración tributaria. Como el número de identificación personal sueco contiene la fecha de nacimiento, también se puede usar para describir la estructura de edad del personal. Si, por ejemplo, 30% del personal tuviera un dígito incorrecto para el mes en su número de registro, eso no afectaría en particular el análisis estadístico, aunque los cálculos salariales e impositivos se volverían imposibles.

1.5.4 ¿Cómo se deben procesar los datos administrativos para fines estadísticos?

Un principio general consiste en combinar muchas fuentes cuando se crea un registro estadístico. Esto se debe a varias razones, entre ellas, el uso de variables de distintas fuentes para enriquecer el contenido —como señalamos en las secciones 1.4.1 y 1.4.2—, además del mejoramiento de las posibilidades de cobertura y validación.

Cuando el conjunto de objetos de una fuente administrativa no es adecuado como población estadística relevante, se deben combinar varias fuentes para crear un conjunto de objetos con una buena cobertura.

Ejemplo: Un registro empresarial en una oficina nacional de estadística se basa en fuentes administrativas. Con cinco fuentes creamos un registro de empresas para Suecia que contenía a todas las empresas (unidades jurídicas) activas en 2002. Cada fuente está compuesta por las unidades jurídicas de un sistema de tributación. En la tabla siguiente, se comparan subcobertura y sobrecobertura de las fuentes utilizadas en nuestro registro de empresas final. La primera fuente contiene la información más antigua disponible y la quinta, la más reciente. Los conjuntos de objetos administrativos de cada fuente son adecuados para cada uno de los cinco sistemas tributarios. Aunque consideradas individualmente las fuentes son de baja calidad *estadística*, combinándolas todas se logra una buena cobertura.

Gráfico 1.8 Sobrecobertura y subcobertura en cinco fuentes administrativas

	Fuente 1	Fuente 2	Fuente 3	Fuente 4	Fuente 5
Sobrecobertura	41%	0%	0%	0%	0%
Subcobertura	21%	74%	74%	30%	9%

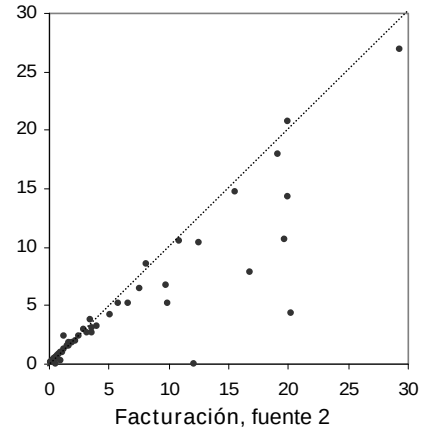
Por lo general, en las encuestas por muestreo y los censos sólo se usan los datos captados por el proyecto para hacer la validación. En cambio, en las encuestas basadas en registros, es posible comparar variables de distintas fuentes, lo que ofrece mayores posibilidades de identificar y corregir errores, por ejemplo, errores de medición y otros tipos de inconsistencias.

Ejemplo: En el gráfico 1.9 se comparan las unidades empresariales (cada una integrada por varias unidades jurídicas) consideradas en una encuesta sueca de empresas respecto a la facturación registrada en dos fuentes administrativas.

A causa de las diferentes definiciones, la fuente 2 debe registrar una mayor facturación que la fuente 1. Esto significa que deben verificarse las cuatro unidades empresariales sobre la línea diagonal porque probablemente estén incompletas.

Asimismo, se deben verificar las unidades empresariales con grandes desviaciones debajo de la línea diagonal, pues probablemente los informes de la fuente 1 se han retrasado para algunas unidades jurídicas pertenecientes a esas unidades empresariales.

Gráfico 1.9 Comparación de fuentes



1.6 ¿POR QUÉ USAR DATOS ADMINISTRATIVOS PARA PRODUCIR ESTADÍSTICAS?

Hay tanto prejuicios como críticas justificadas en relación con las estadísticas basadas en registros administrativos. Uno de los prejuicios es que, si bien de bajo costo, se considera que ese tipo de estadísticas son de mala calidad en comparación con las estadísticas "verdaderas" basadas en encuestas. Las críticas justificadas se refieren a errores de relevancia, problemas de comparabilidad y al hecho de que la calidad estadística no está bajo control. Desafortunadamente, podemos hacer críticas similares a las encuestas por muestreo, donde los errores del marco, los errores de medición y los errores por la no respuesta pueden ser significativos y, pese a ello, pasar inadvertidos, independientemente de que los errores de muestreo estén o no bajo control.

Nuestra respuesta a este tipo de comentarios es que las estadísticas sobre la sociedad deben incluir tanto estadísticas basadas en registros como estadísticas basadas en datos captados por una oficina nacional de estadística. De modo que no se puede decir que un método sea mejor que otro, pues en ciertos casos serán más eficaces las encuestas basadas en registros y en otros las encuestas por muestreo.

1.6.1 Pros y contras de las encuestas basadas en registros

Como dijimos, es común que se piense, a menudo sin un sustento sólido, que las estadísticas basadas en registros son de mala calidad. ¿Se justifica esta idea en el caso de las fuentes administrativas que usa una oficina de estadística? Gran parte del sistema de registros de la Oficina de Estadística de Suecia se basa en datos del registro poblacional administrativo y

de la administración tributaria. ¿Serían de mejor calidad estas estadísticas si la propia Oficina de Estadística de Suecia recabara los datos de manera paralela a la captación de datos administrativos efectuado por el Consejo Nacional Tributario? Difícilmente. Si la Oficina de Estadística de Suecia intentara recopilar esos datos, sería una tarea costosa, aumentaría la carga para los informantes y probablemente se generarían datos con más errores de medición.

Gráfico 1.10 Pros y contras de las encuestas basadas en la captación de datos y en registros administrativos

	Pros	**Contras**

Encuestas basadas en captación de datos: encuestas por muestreo y censos	- Se puede elegir qué preguntas formular.
	- Pueden estar actualizadas.
		- Algunos informantes:
		  - no entienden la pregunta;
		  - han olvidado la respuesta;
		  - no responden (no respuesta);
		  - responden con descuido.
		- La carga para los informantes puede ser pesada.
		- Alto costo.
		- Mala calidad de las estimaciones para campos de estudio pequeños (en el caso de las encuestas por muestreo).

Encuestas basadas en registros administrativos	- No hay una mayor carga para el informante. - Bajo costo. - Cobertura casi completa de la población. - Cobertura de tiempo completa. - Los entrevistados responden con cuidado las preguntas administrativas importantes. - Buenas posibilidades de presentar informes sobre áreas pequeñas, estadísticas regionales y estudios longitudinales.	- No se pueden hacer preguntas. - Dependen de las definiciones del sistema administrativo para la población, el objeto y las variables. - La elaboración de informes sobre datos administrativos puede ser lenta debido al largo tiempo que puede transcurrir entre el periodo de referencia y el momento en que los datos están disponibles para fines estadísticos. - Los cambios en los sistemas administrativos dificultan las comparaciones. - Las variables de menor importancia para el trabajo administrativo pueden ser de menor calidad.

En este cuadro resumimos las ventajas y desventajas de los dos métodos de encuesta. La trascendencia de las desventajas varía dependiendo de cada situación.

Nuestro ejemplo de la transición de los jóvenes de los estudios al mercado de trabajo puede ilustrar tanto las ventajas como las desventajas de las estadísticas basadas en registros. Ya mencionamos que este tipo de encuesta es adecuado para los estudios longitudinales. Otra ventaja radica en que es posible dar resultados para varios subgrupos, diferentes carreras, programas de estudio y regiones. Sin embargo, una desventaja de ciertos tipos de fuentes administrativas es que con los sistemas administrativos a menudo se debe esperar mucho desde el periodo de referencia hasta que los datos están disponibles, lo que puede originar retrasos con las estadísticas basadas en registros. En nuestro ejemplo, el informe correspondiente a 1987–1993 se publicó en junio de 1995, es decir, un año y medio después de finalizado el periodo de referencia.

De modo que no sólo se trata de decidir *si* se usarán datos administrativos, sino también *cómo* se usarán. Nuestra respuesta al *cómo* es que, en términos generales, los datos administrativos no deben usarse tal cual, sino que deben procesarse para fines estadísticos. Se debe intentar construir un sistema de registros coordinados, lo que puede resultar beneficioso en lo que respecta a calidad y costo. Asimismo, el aseguramiento de la calidad debe ser un componente importante del sistema.

Hay muchas tendencias del desarrollo que amenazan la calidad de las estadísticas basadas en encuestas por muestreo o censos. El uso creciente del *telemarketing* o la identificación de llamadas dificulta que se lleven a cabo entrevistas telefónicas. Si la motivación del entrevistado disminuye, aumentarán las no respuestas y los errores de medición. Será más complicado motivar un doble suministro de datos: ¿por qué responder un cuestionario sobre la facturación de la empresa si también se está presentando ante el fisco una declaración del impuesto al valor agregado (IVA) que contiene esa misma información? Todas estas circunstancias hacen que las estadísticas basadas en registros adquieran más importancia. En el siguiente fragmento de una nota periodística tenemos un ejemplo de que el doble suministro de datos a la oficina de estadística y a otra autoridad resulta excesivo:

> Mariestads Tidning a 26 de junio de 2000:
>
> **¡No al envío de datos a la Oficina de Estadística de Suecia!**
> El Sr. R de la granja B piensa que las autoridades deben ser capaces de obtener la información de sus propios registros.
>
> El Sr. R se rehúsa a enviar datos a la Oficina de Estadística de Suecia. Como envía información cada dos semanas al Consejo de Agricultura de Suecia, considera que las autoridades deben cooperar entre sí en vez de…

1.6.2 Costos: ¿las estadísticas basadas en registros son baratas?

"Queda bastante claro que los estados miembros se hallan en la situación paradójica de tener que enfrentar varios recortes presupuestarios al tiempo que ofrecen a los usuarios un volumen mayor de información relevante de alta calidad". Con estas palabras, Yves Franchet, entonces director general de la Oficina Estadística de las Comunidades Europeas (Eurostat), inauguró un seminario (Eurostat, 1997) sobre el uso de las fuentes administrativas para fines estadísticos. La cita ilustra la necesidad de sistemas estadísticos más eficaces, lo cual se puede lograr combinando dos estrategias:

- El mayor uso de fuentes administrativas reduce el costo de la captación de datos tanto para los informantes como para los productores de estadísticas.
- El uso más eficaz y flexible de los datos disponibles permitiría satisfacer nuevas solicitudes evitando el costo de captar nuevos datos. Esto podría lograrse si se creara un sistema de registros estadísticos coordinados a partir de fuentes administrativas.

La construcción de un sistema de registros de ese tipo tendría un costo elevado, pero los costos marginales podrían disminuir porque se le daría un uso más intensivo. En el ejemplo anterior sobre la transición de los jóvenes de la etapa educativa a la vida laboral, el costo marginal del producto fue reducido.

La carga para los informantes también representa un costo que se puede mitigar pasando de la estadística basada en la captación de datos a la estadística basada en registros.

Las encuestas por muestreo son un método cada vez más caro. Además, también están aumentando tanto el número de llamadas telefónicas para establecer el primer contacto

como el costo de llamadas a un porcentaje cada vez mayor de informantes para recordarles que envíen sus cuestionarios contestados de regreso a la oficina de estadística.

Los costos por habitante de las encuestas por muestreo son mayores para los países pequeños —porque si se desea lograr el mismo nivel de precisión en ciertos campos de estudio, se requiere casi el mismo tamaño de muestra en un país con poca población que en uno más grande—. Por lo tanto, para los países pequeños es de especial interés establecer un sistema estadístico basado en registros.

1.7 PANORAMA GENERAL DEL LIBRO

A lo largo de todo el libro presentamos propuestas de cambio. Son necesarios un nuevo enfoque, una nueva terminología y nuevas metodologías para que el sistema de registros y la estadística basada en registros cobren auge y funcionen de una manera más eficiente. A continuación presentamos una síntesis del contenido de cada capítulo.

Capítulo 1 Encuestas basadas en registros: introducción
Para el desarrollo de la estadística basada en registros, se requiere una *teoría estadística para registros*, con sistemas estadísticos como parte medular. Un sistema de registros funcional constituye la base de una producción de estadísticas eficiente.

Capítulo 2 Cómo estructurar un sistema de registros
Presentamos la estructura del sistema de registros. El *modelo de registro* es una importante herramienta para contribuir a una comprensión más amplia del sistema de registros. Hablamos de la importancia de las *fuentes administrativas* y definimos el papel de los cuatro *registros básicos*.

Capítulo 3 Una terminología para encuestas basadas en registros
Introducimos varios *términos estadísticos para registros*. Una terminología adecuada es indispensable para evitar una teoría de los registros imprecisa y facilitar el intercambio de experiencias. Proponemos términos para diferentes tipos de registros y definimos las variables según su función en el sistema de registros. También describimos las variables derivadas mediante emparejamiento.

Capítulo 4 Encuestas por muestreo y registros
¿Cuáles son los beneficios del sistema de registros para las encuestas por muestreo? y ¿cómo combinar encuestas por muestreo y registros? Presentamos un esquema de estos temas y analizamos las diferencias entre los métodos aplicados en las encuestas por muestreo y las encuestas basadas en registros. Es importante distinguir entre encuestas según su metodología: por un lado, las encuestas por muestreo y los censos y, por el otro, las encuestas basadas en registros. En las encuestas con una captación de datos propia, la fase de captación de datos corresponde a los diferentes tipos de procesamiento de registros para crear un registro estadístico. Este procesamiento de registros se debe estudiar desde la perspectiva de las ciencias estadísticas.

Capítulo 5 Cómo crear un registro: la población
El procedimiento para crear un registro consta de cinco fases. En las estadísticas basadas en registros, *población de registro* es un término fundamental que corresponde al término *marco de población* en las encuestas por muestreo. Se deben usar todas las fuentes disponibles para crear una población de registro con buena cobertura. Al combinar diferentes fuentes administrativas, también es posible definir una población de registro

estadísticamente relevante. Presentamos una serie de principios, como *"todos deben usar las poblaciones estandarizadas de los registros base"* y *"todos deben apoyar los registros base para que éstos tengan la mayor calidad posible"*.

Capítulo 6 Cómo crear un registro: las variables

Las variables derivadas son cruciales para las estadísticas basadas en registros. Las variables administrativas se usan para definir variables derivadas estadísticamente relevantes. Examinamos la validación de datos administrativos apoyándonos en varios estudios de caso. Hay algunas diferencias importantes entre la validación de datos en una encuesta por muestreo y en un registro administrativo, pues en este último caso se valida la consistencia de muchas fuentes al mismo tiempo y no sólo se buscan errores en las variables, sino también en los objetos.

Capítulo 7 Métodos de estimación

Los *métodos de estimación* también son necesarios en las estadísticas basadas en registros. No siempre basta con sumar los datos. El enfoque tradicional dentro de las estadísticas basadas en registros es que no se requieren métodos estadísticos. Sin embargo, el mismo enfoque estadístico que aplican actualmente quienes trabajan con problemas de estimación para encuestas por muestreo y censos se debe aplicar también a las estadísticas basadas en registros. Distinguimos entre métodos de estimación básicos y suplementarios. Los métodos suplementarios usan ponderadores de manera similar a las encuestas por muestreo. Ejemplificamos el uso de los ponderadores y su calibración.

Capítulo 8 Calibración e imputación

Se pueden formular distintos métodos de estimación para abordar problemas derivados de la *no respuesta* o *valores perdidos*, *sobrecobertura* y *cambios de nivel* en las series de tiempo. Cuando se integran diferentes registros y se importan variables de un registro a otros, también se importan los defectos de calidad, por ejemplo, los valores perdidos. Esto significa que no basta con hacer ajustes por valores perdidos en una variable en un solo registro, sino que se deben ajustar los valores perdidos para esa variable en todo el sistema de registros de una manera consistente.

La sobrecobertura puede originar errores graves en las estadísticas basadas en registros. Proponemos la calibración de ponderadores para hacer los ajustes necesarios en caso de sobrecobertura en el registro base. Luego se aplicarán esos ponderadores al resto de los productos estadísticos que usen el registro base. La calibración de ponderadores también se puede usar para ajustar los cambios de nivel en las series de tiempo en el nivel micro. Con estos ponderadores, se pueden producir series de tiempo consistentes y vinculadas.

Capítulo 9 Estimación con objetos combinados

Las variables de valores múltiples, como "rama industrial", se usan de tal modo que se originan errores de agregación. En este capítulo presentamos métodos de estimación especiales basados en ponderadores que pueden ser útiles para reducir los valores de agregación en estas variables. Asimismo, introducimos los objetos combinados para los problemas de estimación relacionados con las variables de valores múltiples. Estos objetos también pueden servir para ajustar los cambios de nivel de las series de tiempo en el nivel micro.

Capítulo 10 Calidad de las estadísticas basadas en registros

Por largo tiempo se ha considerado que los errores de muestreo son el error más importante en las encuestas por muestreo. En las encuestas basadas en registros no hay fase de muestreo, sino que predomina la *fase de integración*, donde los datos de distintas fuentes se integran en un nuevo registro estadístico.

Durante la fase de integración, se generan la población de registro y los objetos derivados, se importan variables de distintas fuentes y se crean variables derivadas. Los tipos de errores que se originan en la fase de integración deben recibir el nombre de *errores de integración*. Se requieren *indicadores de calidad* específicos que se adapten a las necesidades de las estadísticas basadas en registros. Presentamos un conjunto de indicadores, entre los cuales se deben seleccionar los pertinentes para un registro específico. Además, se debe llevar a cabo una *evaluación general* de la calidad.

Capítulo 11 Metadatos y sistemas informáticos

Las estadísticas basadas en registros requieren un sistema de metadatos en el que puedan procesarse grandes cantidades de *metadatos formalizados* usando las herramientas informáticas apropiadas. Son componentes importantes de este sistema un *calendario* con los cambios importantes y una *base de datos de definiciones.*

La documentación del sistema informático para estadísticas basadas en registros debe ser distinta a la *documentación de sistemas* tradicional, que relaciona las encuestas con su captación de datos propia. *La tecnología de almacenamiento de datos* puede ser una herramienta para:

– una gestión más eficaz del registro;
– una mayor cantidad de metadatos y metadatos más uniformes;
– recuperaciones más simples y seguras;
– un mejor panorama general del contenido del sistema.

Capítulo 12 Protección de la privacidad y la confidencialidad

Las rutinas eficaces para la *protección de la privacidad* son un elemento muy importante del sistema de registros. Se debe considerar el uso mínimo de variables con información en lenguaje llano y números de identificación oficiales. Siempre se deben incluir rutinas de *protección contra la revelación de datos confidenciales* cuando se presentan y difunden tablas basadas en registros estadísticos y cuando se dan a conocer microdatos para fines de investigación.

Capítulo 13 Coordinación y coherencia

El concepto *coherencia* se refiere a que se pueden usar en conjunto estimaciones de diferentes encuestas. Por ejemplo, para que un cociente sea significativo, el numerador y el denominador deben ser coherentes. Las encuestas consistentes arrojan estimaciones coherentes.

Las estadísticas de fuentes diversas serán consistentes, es decir, tendrán un alto grado de coherencia si:

– se asegura la consistencia respecto a las *poblaciones* (en relación con las definiciones de objeto y conjunto de objetos);
– se asegura la consistencia respecto a las *variables*;
– se usan métodos de calibración que arrojen *estimaciones* consistentes.

Mediante un ejemplo basado en empresas ilustramos un método para lograr estadísticas coordinadas y consistentes. El propósito es demostrar que puede haber inconsistencias en encuestas realizadas en diferentes momentos.

Capítulo 14 Conclusiones

En el último capítulo presentamos algunas conclusiones generales de los capítulos anteriores. Es necesaria una nueva manera de abordar los datos administrativos y se debe reconocer que las estadísticas basadas en registros representan un importante campo para las ciencias estadísticas.

CAPÍTULO 2

Cómo estructurar un sistema de registros

En el capítulo 1 se menciona varias veces el término *sistema de registros*. En las secciones 1.4.1 y 1.4.2 se presentan ejemplos que ilustran la importancia del sistema de registros en la creación de nuevos registros estadísticos. En el gráfico 1.1 señalamos que todos los registros estadísticos deben estar incluidos en un sistema de registros coordinado. Este sistema asegurará que todos los datos se puedan integrar y usar de manera eficiente. En este capítulo, describimos el sistema de registros estadísticos. Un productor de estadísticas necesita un modelo para describir los registros que existen en la organización y los vínculos entre ellos.

En conjunto se puede decir que todos los registros son un sistema de información estadística. Resulta más fácil entender este sistema mediante un modelo conceptual, que sin entrar en demasiados detalles técnicos describe los tipos de objeto del sistema y las relaciones entre ellos. Los modelos son necesarios, por un lado, para describir como es el sistema en realidad y, por el otro, para delinear una mejor estructura que sirva de punto de partida para el desarrollo del sistema. En este capítulo nos referiremos al sistema de registros actual de la Oficina de Estadística de Suecia para explicar un modelo de registro, cuya estructura es de interés general incluso para oficinas de estadística con sistemas menos avanzados.

En primer lugar, explicaremos la estructura general del sistema. Después describiremos los *registros base* del sistema, presentando como ejemplos los registros de la Oficina de Estadística de Suecia. Tras un análisis de las *variables estandarizadas*, veremos otros ejemplos de sistemas de registros estadísticos fuera de dicha dependencia.

2.1 UN MODELO DE REGISTRO BASADO EN TIPOS Y RELACIONES DE OBJETOS

Para empezar, debemos establecer una distinción entre los términos *registro* y *encuesta basada en registros*. Un registro específico puede usarse en diferentes encuestas basadas en registros; por tanto, es posible hablar de registros sin mencionar qué encuesta en particular los usará. Así es como nos referiremos a los registros y el sistema de registros en este capítulo. Utilizaremos términos informáticos y sólo recurriremos a la terminología estadística cuando hablemos de *encuestas* basadas en registros. En el gráfico 2.1 se muestra la relación entre los términos informáticos y los estadísticos.

Gráfico 2.1 Relación entre términos informáticos y estadísticos

Términos informáticos	Términos estadísticos
Objeto	Objeto, unidad estadística, elemento
Tipo o clase de objeto	Tipo de unidad estadística
Conjunto de objetos	Población
Relaciones entre objetos	Vínculos

El primer paso para diseñar este modelo conceptual del sistema de registros fue elaborar un inventario de todos los registros y productos de registros que había en la Oficina de Estadística de Suecia. Una encuesta por muestreo comienza por la definición de los objetos o unidades estadísticas de la población. De modo que empezamos a trabajar con el modelo conceptual clasificando los registros del inventario por tipo de objeto. Sólo estudiamos tipos de objetos con una importancia estadística, es decir, aquellos incluidos en las poblaciones objetivo. En algunos registros hay jerarquías de tipos de objetos, por lo que resulta natural agruparlos de la siguiente manera:

– persona y hogar
– organización, empresa y establecimiento donde una empresa lleva a cabo sus actividades
– inmueble o propiedad, edificio y vivienda
– vehículo

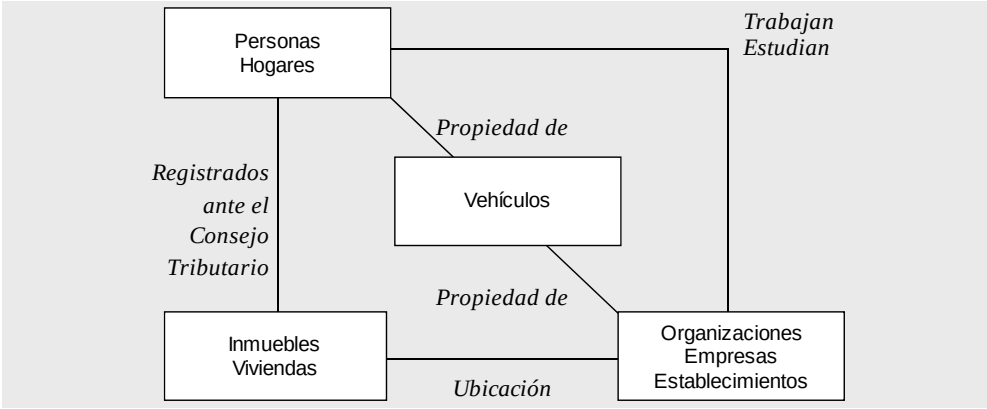
Puede haber *relaciones* entre diferentes tipos de objetos:

– Entre una persona, una empresa y un establecimiento: una persona *está empleada* por una empresa u organización y *trabaja* en un establecimiento específico. Una relación similar se aplica entre estudiantes y lugares de estudio: una persona *participa en un programa de educación* de una organización o empresa específica y la enseñanza se lleva a cabo en cierto lugar de estudio o establecimiento.
– Entre una persona y una propiedad o vivienda: una persona *está registrada* (ante las autoridades hacendarias) en una propiedad o vivienda específica.
– Entre establecimiento y propiedad: un establecimiento *se encuentra* en un domicilio específico y el domicilio se relaciona con una propiedad específica.
– Entre una persona y un vehículo, y entre una empresa y un vehículo: una persona o una empresa *es propietaria* de un vehículo específico.

Estas relaciones son muy importantes desde el punto de vista administrativo y jurídico. Por ello, las fuentes administrativas contienen información de muy buena calidad sobre estas relaciones. También son fundamentales para un sistema de registros funcional: cuando una fuente administrativa contiene información sobre relaciones, se crean *vínculos* en el sistema.

Un vínculo entre dos objetos consiste en una o varias *variables de vinculación* comunes, que contienen la información necesaria para identificar relaciones entre objetos. Como las relaciones antes mencionadas se refieren a distintos tipos de objetos, esos vínculos tienen una función relevante en el sistema de registros: en el modelo final son los vínculos entre los registros base del sistema. Creamos el modelo siguiente con base en esos objetos y relaciones. Los rectángulos representan *conjuntos de objetos* y las líneas, *relaciones* entre objetos.

Gráfico 2.2 Modelo conceptual del sistema de registros, versión 1



En el gráfico 2.2 vemos cuatro registros que contienen diferentes tipos de objeto de interés estadístico específico. Cada rectángulo corresponde a uno de los cuatro registros de la Oficina de Estadística de Suecia: *población, vehículos, inmuebles y empresas*.

Cuando hablamos del Registro de Población, en realidad nos referimos a una serie de registros porque ese registro existe en muchas versiones: uno de personas, otro de familias, otros de la población en diferentes momentos, etcétera.

La relación *Registrado ante el Consejo Tributario* se muestra en el Registro de Población mediante una variable de vinculación que permite identificar la propiedad inmobiliaria donde el Consejo Tributario tiene registrada a una persona específica. La relación *Propiedad de* se muestra en el Registro Vehicular mediante la inclusión de la identidad personal o de organización del propietario de un vehículo como una variable de vinculación.

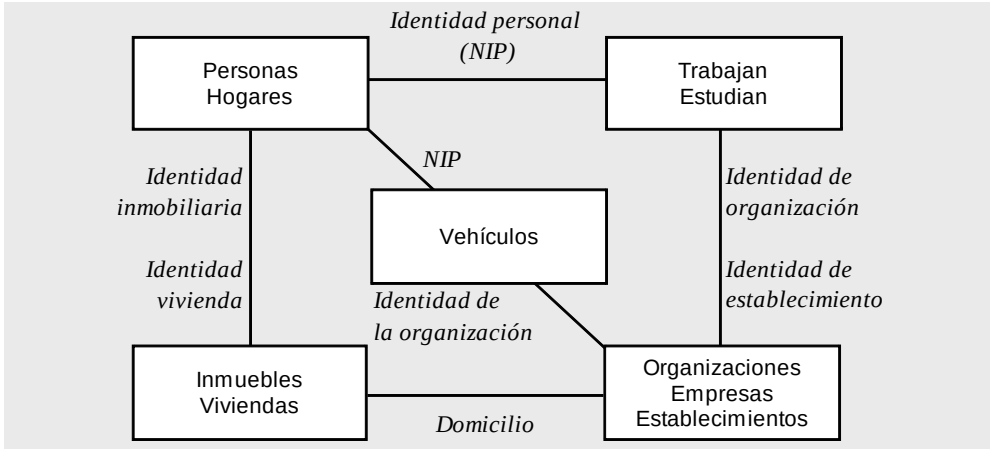
Ni el Registro de Población ni el Registro Empresarial contienen variables de vinculación que muestren la relación *Trabajan/Estudian*, pero hay otros registros con información sobre empleo remunerado o actividades educativas:

– El Registro de Estados de Resultados (un registro de empleo basado en los pagos de impuestos anticipados) y los registros salariales contienen variables que proporcionan información sobre empleo. El tipo de objeto *empleo* está definido por una relación entre los tipos de objeto *trabajador*, *empleador* y *establecimiento*. Por cada combinación de identidad personal, identidad empresarial o de organización, e identidad del establecimiento, hay datos sobre ingresos o salario, etcétera.
– Los registros con diferentes tipos de estudiantes contienen variables que muestran la relación entre personas que estudian, organizadores de cursos y lugares de estudio. Los organizadores de cursos son organizaciones o empresas y los lugares de estudio, establecimientos.

¿Cómo se debe representar la información sobre la relación *Trabajan/Estudian* en el sistema? Una relación entre dos tipos de objeto se puede considerar un tipo de objeto: un *objeto relacional*. Cuando hay *muchas* variables que describen estas relaciones, conviene más tomarlas como objetos. Como hay muchas variables de interés estadístico en las diferentes fuentes que describen *Trabajan/Estudian*, la mejor solución radica en crear un registro especial formado por objetos relacionales, es decir, objetos que describan la relación entre personas, empresas y establecimientos, donde la misma persona puede tener varias relaciones con distintas empresas y establecimientos. En las secciones 2.2.2, 2.2.5 y 3.3.4 veremos con más detalle a qué se debe esto.

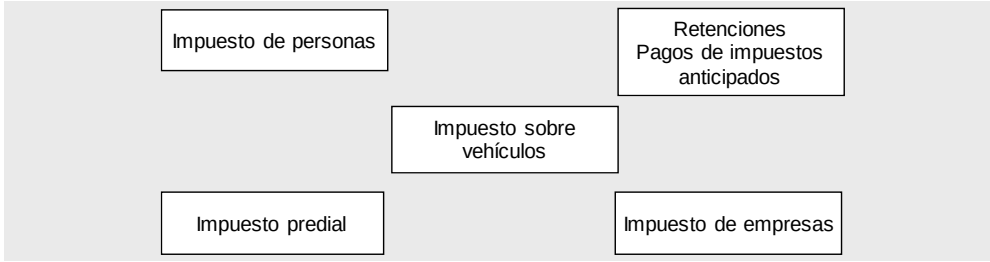
En el gráfico 2.3 introducimos el objeto relacional *Trabajan/Estudian* como un registro especial al que llamamos *Registro de Actividades*. Las relaciones del gráfico 2.2 han sido remplazadas por las variables que funcionan como vínculos entre objetos en los diferentes registros.

Gráfico 2.3 Modelo conceptual del sistema de registros, versión 2



El sistema de registros estadísticos de la Oficina de Estadística de Suecia se basa en distintas fuentes administrativas. En el siguiente gráfico vemos los cinco componentes del sistema tributario que son fuentes de los cinco registros mostrados en el gráfico 2.3.

Gráfico 2.4 El sistema de registros se basa en distintos componentes del sistema tributario



Además de los cinco registros mostrados en el gráfico 2.3, hay aproximadamente 50 registros más en la Oficina de Estadística de Suecia. Con base en las variables identificadoras existentes, se pueden establecer vínculos con los registros de Población, Actividad, Inmobiliario o Empresarial. Estos cuatro registros desempeñan un papel de particular importancia dentro del sistema de registros, por lo que reciben el nombre de *registros base*. En el gráfico 2.5 se muestra todo el sistema de registros en el modelo conceptual general.

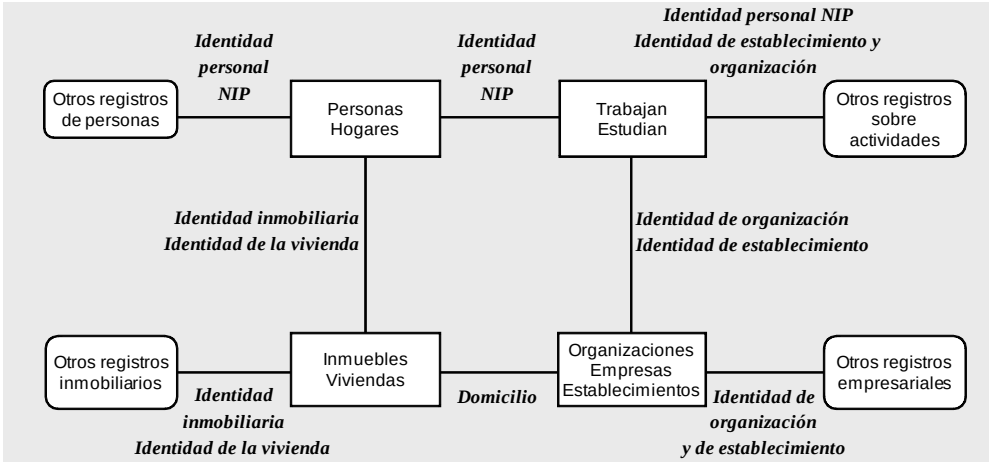
El Registro Vehicular contiene el tipo de objeto *vehículo*, que tiene un papel más limitado dentro del sistema. Para fines de simplificación, incluimos vehículos propiedad de personas en los diversos registros sobre personas y vehículos propiedad de empresas en los diversos registros sobre empresas.

El modelo final se presenta a continuación. ¿Este modelo se puede usar de manera generalizada? Como en muchos países existen los mismos tipos de objeto y contenidos de variables, consideramos que este modelo es adecuado en términos generales para describir sistemas de estadísticas sobre la sociedad. Una oficina nacional de estadística, con acceso a los

registros estadísticos mencionados en el gráfico 2.4, puede construir un sistema de registros estadísticos de acuerdo con la estructura mostrada en el modelo final del gráfico 2.5.

En los países escandinavos, el desarrollo de las estadísticas basadas en registros se inició con los registros sobre personas. En otros países, es más fácil tener acceso a los registros empresariales administrativos, por lo que el desarrollo comienza en esa parte del sistema. Sin embargo, el modelo final puede ser el mismo en todos los países.

Gráfico 2.5 Modelo conceptual de un sistema de estadísticas basado en registros sobre la sociedad



2.2 EL SISTEMA DE REGISTROS BASE

Los cuatro registros base, y los vínculos entre ellos, constituyen la columna vertebral del sistema de registros pues contienen los tipos de objetos y los vínculos importantes. Si alguno de estos registros base falta o es de mala calidad, todo el sistema de registros será de mucho menor utilidad para fines estadísticos. Los registros base tienen como función definir los objetos y poblaciones en el sistema, de modo que una buena definición de los objetos y una buena cobertura son características fundamentales para garantizar la calidad de todo el sistema.

Las variables importantes en un registro base son aquellas que identifican un objeto y se pueden usar para vincular ese objeto con los objetos de otros registros. La información sobre tiempo relativa a los distintos eventos que afectan al objeto también es necesaria para crear poblaciones relacionadas con un momento o periodo específicos.

Además de estar sustentado en fuentes administrativas estables y confiables, un registro base debe poseer las siguientes características:

Gráfico 2.6 Características de un registro base

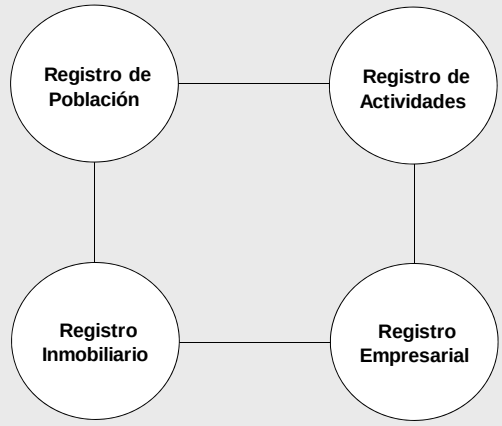
1. Define tipos de objeto importantes.
2. Define conjuntos de objetos o poblaciones estandarizadas importantes.
3. Contiene vínculos con objetos de otros registros base.
4. Contiene vínculos con otros registros relacionados con el mismo tipo de objeto.
5. Es importante para el sistema en su conjunto, por lo que resulta esencial que sea de alta calidad y esté bien documentado.

6. Es importante para el marco muestral.
7. Se puede usar para estadísticas demográficas relacionadas con personas, actividades, inmuebles o empresas.

   Así como se describe la distribución por edad, los nacimientos y las defunciones en una población de personas, debe ser posible describir la distribución por edad, los nacimientos y las defunciones en empleos, edificios o establecimientos. Las fechas de nacimiento y defunción deben estar disponibles en el registro base para producir estadísticas demográficas.

En el gráfico 2.7 se describen los cuatro registros base.

Gráfico 2.7 Los cuatro registros base



El Registro de Población de este gráfico se refiere a múltiples versiones del registro porque cada registro base puede existir en varias versiones:

- población actual;
- población en un momento específico, por ejemplo, 31 de diciembre;
- todos los cambios durante un periodo específico, por ejemplo, año calendario;
- todos los objetos que han existido en un periodo específico, por ejemplo, el registro por año calendario.

Ahora examinaremos estos cuatro registros base y la forma en que los registros existentes en la Oficina de Estadística de Suecia cumplen con los siete requisitos mencionados.

2.2.1 Registro de Población

Este registro base es el más avanzado en cuanto a la formulación de métodos estadísticos para registros aplicables a un registro base. Cumple con los siete requisitos señalados en el gráfico 2.6. Por consiguiente, puede servir como modelo para otros registros base. Quienes trabajan con el Registro de Población saben cómo actualizarlo, tienen una larga experiencia en la producción de estadísticas avanzadas basadas en registros, regularmente realizan análisis demográficos y están acostumbrados a generar registros integrados para diferentes propósitos. Se ha formulado un sistema de producción avanzado en el que se almacena el historial de los objetos.

El registro se basa en datos administrativos del sistema de registros cívicos del Consejo Nacional Tributario (CNT). La Oficina de Estadística de Suecia capta estos datos diariamente por medio del sistema del CNT y también recibe algunos datos anuales del Consejo de Migración sueco. Los datos se refieren a las variables de identificación y ubicación para personas y a variables demográficas como edad, sexo, estado civil, etcétera.

Cuando ocurrió la transición de las computadoras centrales a las redes de computadoras personales a finales de los años noventa, aumentó el contenido de variables del Registro de Población. La mejor calidad de los datos sobre *domicilio*, *relaciones* con otras personas y *fechas* para todas las ocurrencias en el Registro de Población ofrece nuevas posibilidades para la producción estadística. El emparejamiento se facilita mediante la consignación de números de identificación personal corregidos y temporales. Además, los diferentes tipos

de correcciones se consignan de modo que sea posible diferenciar entre los eventos reales y otras correcciones. Estas mejoras son importantes en lo que respecta a la función del registro como registro base. Wilén y Johannesson (2002) describieron el nuevo Registro de Población.

No obstante lo anterior, hay algunos aspectos negativos:

– La definición de hogares presenta ciertas lagunas que podrían remediarse mediante el establecimiento de un buen registro de viviendas y si las personas estuvieran registradas por vivienda como en Dinamarca y Finlandia.
– Hay sobrecobertura para ciertas categorías de inmigrantes. La Oficina de Estadística de Suecia descubrió esto y lo informó al CNT, que desde entonces ha verificado con mayor atención el Registro de Población.

2.2.2 Registro de Actividades

Este registro consta de tres partes: *empleo o actividades laborales remuneradas*, *actividades de estudio* y *otras actividades relacionadas con el mercado laboral*. Como ya se mencionó, el registro consiste en objetos relacionales. Cada uno de esos objetos se identifica mediante tres variables de vinculación: número de identificación personal, número de organización y número de establecimiento del Registro Empresarial.

Los datos sobre *actividades laborales* se basan en el estado de resultados anual del empleador, que contiene el ingreso para cualquier combinación de *empleado* y *establecimiento*. Las declaraciones de impuestos aportan información sobre las personas que están activas como *trabajadores por cuenta propia*. En el caso de los empleados del sector público, la Oficina de Estadística de Suecia recibe información mensual de la nómina de los empleadores.

Las siguientes son características de las *actividades laborales* dentro del Registro de Actividades: ingreso percibido, ingreso mensual, duración del trabajo, ocupación y transporte entre el lugar de residencia y el lugar de trabajo.

Los datos sobre *actividades de estudio* se basan en datos de las escuelas y universidades. Hay diversos registros para diferentes tipos de estudiantes. Estos registros contienen los números de identificación personal de los estudiantes, así como detalles sobre la escuela o lugar de estudio, que a su vez es un establecimiento en el Registro Empresarial. Al asignar un número de establecimiento a estos lugares de estudio es posible vincular el Registro Escolar con el Registro Empresarial. La actividad *estudiando* en el Registro de Actividades será entonces una relación entre un número de identificación personal y un número de establecimiento. El transporte entre el lugar de residencia y el lugar de estudio se puede identificar para los estudiantes de la misma manera que se hace para las personas con un empleo remunerado: con un número de identificación personal y un número de establecimiento que indiquen la ubicación del lugar de residencia y del establecimiento.

Los datos sobre *otras actividades relacionadas con el mercado laboral* también podrían incluirse en el nuevo registro. Diferentes autoridades tienen información sobre servicio militar, subsidios por enfermedad, pensiones por discapacidad, políticas de empleo, desempleo registrado y atención médica institucional. Esta información puede ofrecer un panorama complementario sobre el estado del mercado laboral además de información sobre empleo remunerado y estudios. Las fuentes administrativas contienen información que permite ubicar actividades en cuanto a tiempo, aunque su calidad no siempre es la mejor.

Hay una gran cantidad de datos administrativos que contienen vínculos tanto con personas como con empresas. Desde una perspectiva puramente técnica, podría considerarse que

estos datos describen relaciones entre personas y empresas. Como dijimos antes, una relación entre dos tipos de objeto puede verse como un objeto relacional. Como hay muchas variables estadísticas que describen estas relaciones, conviene más tomarlas como objetos, que de suyo conllevan un interés estadístico.

Hemos optado por considerar esta información como datos que describen *actividades*. Vemos estas actividades como sus propios objetos por dos motivos, en primer lugar, porque son un tipo de objeto con interés estadístico y, en segundo, porque este tipo de objeto necesita ser distinguible para que el sistema de registros en su conjunto tenga una estructura clara. ¿Satisface el Registro de Actividades los siete requisitos de un registro base mencionados en el gráfico 2.6?

1, 2, 7. Define tipos de objetos, conjuntos de objetos y poblaciones importantes. Puede usarse para estadísticas demográficas en relación con las actividades.
Cuando se cruzan la oferta y la demanda del mercado de trabajo, se crean relaciones entre personas y empresas u organizaciones. Estas relaciones son importantes para las estadísticas del mercado de trabajo y están descritas por muchas variables estadísticas. Conviene tomar estas relaciones como objetos. Las estadísticas demográficas que describen, por ejemplo, la manera en que las diversas actividades remuneradas cambian a raíz de la *creación de empleos* y la *destrucción de empleos*, son muy relevantes para el estudio de las estadísticas del mercado de trabajo. Por tanto, el Registro de Actividades debe contener fechas de creación y extinción de las actividades.

3. Tiene vínculos con objetos de otros registros base.
La variable de vinculación *número de identificación personal* es un vínculo con el Registro de Población y las variables de vinculación *número de organización* y *número de establecimiento* son vínculos con el Registro Empresarial.

4. Tiene vínculos con otros registros que se refieren al mismo tipo de objeto.
Las variables de vinculación *número de identificación personal, número de organización* y *número de establecimiento* son vínculos con el Registro de Estados de Resultados, el Registro Salarial[1] y el Registro Ocupacional. Se estandarizarán otros registros salariales y de estudiantes en lo que respecta a las variables que identifican el establecimiento y el lugar de estudio. El Registro de Actividades tendrá entonces vínculos con todos los registros que contienen estos objetos relacionales.

5. Es importante para el sistema en su conjunto.
Entre los usuarios hay un gran interés en las estadísticas que combinan datos sobre personas y sobre empresas. El Registro de Actividades desempeña un papel muy importante como puente entre estos dos tipos de estadísticas.

6. Es importante como marco muestral.
Podrían seleccionarse ciertas categorías de personas o empresas mediante el Registro de Actividades. También pueden seleccionarse categorías especiales de estudiantes de un registro de actividades estudiantiles.

El Registro de Actividades puede usarse para crear registros sobre personas o empresas
El Registro de Actividades se basa directamente en fuentes administrativas —por cada combinación de persona y establecimiento de una empresa, hay datos administrativos sobre salarios brutos anuales—. Se puede crear un registro sobre personas sintetizando los datos salariales para cada persona y un registro de establecimientos con datos sobre salarios

[1] Usamos el término salarios para referirnos a salarios y sueldos brutos.

brutos sintetizando los datos salariales incluidos en el Registro de Actividades para cada establecimiento ("suma salarial" en el gráfico 2.8).

Gráfico 2.8 Relación entre registros sobre personas, actividades y empresas

Registro de Población: Personas

Persona	Suma salarial
NIP1	450 000
NIP2	210 000
NIP3	270 000

Registro de Actividades: Empleos

Empleo	Persona	Establecimiento	Suma salarial
Em1	NIP1	Es1	220 000
Em2	NIP3	Es1	180 000
Em3	NIP1	Es2	230 000
Em4	NIP2	Es2	210 000
Em5	NIP3	Es2	90 000

El Registro de Actividades contiene la distribución bivariada y los registros de empresas y población contienen distribuciones marginales.

Registro Empresarial: Establecimientos

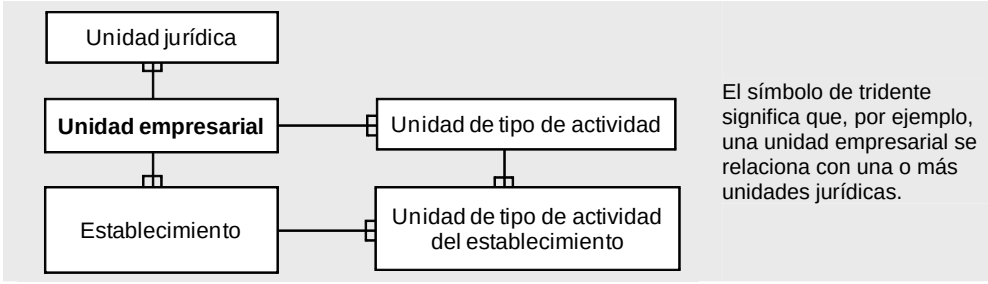
Establecimiento	Suma salarial
Es1	400 000
Es2	530 000

En el gráfico anterior se describen tres registros estadísticos, que ofrecen tres panoramas diferentes pero consistentes de la sociedad. Para entender la estructura del registro, debemos distinguir entre estos tres registros y entender cómo se relacionan. El Registro de Actividades que aparece en el gráfico 2.8 no es un registro sobre personas ni un registro empresarial, sino un registro descriptivo de las relaciones entre personas y empresas. Retomaremos este ejemplo en la sección 3.3.4.

2.2.3 Registro Empresarial

La Oficina de Patentes y Registros y el Consejo Nacional Tributario envían regularmente a la Oficina de Estadística de Suecia datos administrativos concernientes a las *unidades jurídicas*. Se ha llevado a cabo un amplio trabajo en la Oficina de Estadística de Suecia para captar información de empresas con actividades en más de un establecimiento, a fin de elaborar un registro de todos los *establecimientos*. Dentro del Registro Empresarial, también se crean los tipos de objeto *empresas*, *unidades de tipo de actividad* y *unidades de tipo de actividad del establecimiento*.

Gráfico 2.9 Tipos de objeto del Registro Empresarial



Un gran número de fuentes administrativas, usadas por la Oficina de Estadística de Suecia, contienen datos sobre unidades jurídicas. La Oficina de Estadística de Suecia capta datos sobre otros tipos de objetos. Los establecimientos y las unidades de tipo de actividad del establecimiento son importantes para las estadísticas regionales, mientras que las unidades de tipo de actividad y las unidades tipo de actividad del establecimiento lo son para las estadísticas económicas que describen ramas industriales. Las unidades empresariales son

cruciales porque los datos de todas las fuentes, administrativas y estadísticas, se pueden agregar y comparar para estas unidades.

Hay dos principios centrales para asegurar la consistencia[2] interna de las estadísticas económicas:

– En las estadísticas económicas no deben aparecer conjuntos de objetos paralelos. Por ejemplo, la población de productores de energía debe ser la misma en la Encuesta sobre Energía, la Encuesta sobre Estructura Empresarial y el Registro Empresarial.
– En el Departamento de Estadísticas Económicas, todos deben contribuir al mantenimiento del Registro Empresarial. Por ejemplo, quienes captan datos de la industria manufacturera deben remitir la información que obtienen sobre esta población al Registro Empresarial.

Estos principios se deben aplicar a todos los registros base, no sólo al Registro Empresarial. Todos deben participar en el mantenimiento de un registro base, independientemente del departamento en el que trabajen. En la práctica, esto significa, por ejemplo, que quienes trabajan en el Registro Agropecuario deben suministrar información básica sobre el segmento agropecuario del Registro Empresarial y quienes trabajan en el Registro Escolar deben dar la información necesaria para que la sección educativa del Registro Empresarial sea lo mejor posible. Estos principios se abordan con detalle en el capítulo 5.

¿El Registro Empresarial cumple con los siete requisitos de un registro base mencionados en el gráfico 2.6?

1, 2, 7. Define tipos de objetos, conjuntos de objetos y poblaciones importantes. Se puede usar para estadísticas demográficas sobre establecimientos y empresas.
El Registro Empresarial contiene tipos de objetos y poblaciones importantes para estadísticas económicas. Muchos usuarios solicitan estadísticas que describan la demografía de empresas y establecimientos. Johansson (1997, 2001) incluye ejemplos de la manera en que se puede usar el Registro Empresarial para las estadísticas empresariales demográficas dentro de la investigación económica. A continuación presentamos un ejemplo sobre la importancia de mostrar los cambios en la población empresarial. Las estadísticas del Registro Empresarial se publicaron en la primera plana del principal diario sueco:

Dagens Nyheter, julio 12 de 2002

Aquí están los nuevos empleos

¿Le gustaría encontrar empleo? Presente su solicitud en una empresa de servicios. Son ahora las principales generadoras de empleo de acuerdo con las cifras que la Oficina de Estadística de Suecia presenta para Dagens Nyheter…

Debemos ser cautelosos al interpretar las cifras de la Oficina de Estadística de Suecia. Las empresas que parecen haber tenido un mayor crecimiento a menudo son subsidiarias de un grupo corporativo mayor, donde se han reestructurado las operaciones de alguna manera…

Este ejemplo muestra que las estadísticas basadas en el Registro Empresarial pueden tener un valor noticioso, es decir, ser de interés general para un público amplio. El artículo se refiere a la demografía de las empresas. También se menciona que las estadísticas se deben interpretar con cautela, lo cual es necesario porque en este caso los datos administrativos

[2] Usamos el término consistencia no en el sentido de la teoría de inferencias, sino para explicar que diferentes fuentes arrojan estimaciones estadísticas concordantes entre sí.

que recibió la Oficina de Estadística de Suecia no fueron transformados en datos estadísticos. El procesamiento del Registro Empresarial se debe adaptar para satisfacer los requisitos de la estadística —un crecimiento interesante desde el punto de vista estadístico no se debe mezclar con cambios administrativos carentes de interés como fusiones y adquisiciones, como ocurre en el ejemplo anterior—. En la Oficina de Estadística de Suecia, el Registro Empresarial se ha usado sobre todo como marco muestral. Sin embargo, los marcos muestrales no son adecuados para la estadística basada en registros. La demografía de las empresas es un campo importante que en años recientes ha tenido un fuerte desarrollo encaminado a la descripción de los cambios con una relevancia estadística en las poblaciones de empresas. Para satisfacer esas necesidades se debe integrar una versión diferente del Registro Empresarial. Analizaremos este tema en los capítulos 5 y 13.

3. Tiene vínculos con objetos en otros registros base.
La combinación del número de organización de las variables de vinculación y el número de establecimiento es un vínculo con el Registro de Actividades, mientras que el domicilio de ubicación de un establecimiento es el vínculo con el Registro Inmobiliario.

4. Tiene vínculos con otros registros que se refieren al mismo tipo de objeto.
Las unidades jurídicas se identifican mediante números de organización; los otros tipos de objetos del gráfico 2.9 también se identifican con códigos numéricos, que se pueden usar como variables de vinculación. El Consejo Nacional Tributario crea los números de organización, usados en todas las fuentes administrativas, mientras que la Oficina de Estadística de Suecia crea los números de identificación de los otros tipos de objetos. En el caso de los trabajadores autónomos, el número de identificación personal del propietario de la empresa funciona como número de organización. Todas estas variables de vinculación son vínculos con otros registros relacionados con empresas u organizaciones y establecimientos.

5. Es importante para el sistema en su conjunto.
El Registro Empresarial no sólo es importante para las estadísticas de las empresas, sino para las del mercado laboral y las personas. Muchos productos estadísticos usan datos de la rama industrial y la ubicación de un establecimiento.

6. Es importante como marco de muestreo.
Se usa como marco para las encuestas por muestreo relativas a empresas, unidades jurídicas o establecimientos.

2.2.4 Registro Inmobiliario

Se integrará un registro base puro para bienes inmuebles y tipos de objetos relacionados. Este registro se debe actualizar constantemente de modo que se dé seguimiento a las divisiones y fusiones de inmuebles. El registro debe satisfacer los requisitos mencionados en el gráfico 2.6 para un registro base e incluir los siguientes tipos de objetos:

– bien inmueble: tierras;
– bien inmueble: edificaciones y viviendas;
– unidades tributarias y unidades de avalúo.

Las variables importantes en un registro base de este tipo serían las identidades de los diferentes objetos pertenecientes a los tipos de objetos enumerados. También se deben incluir los domicilios de ubicación, las coordenadas geográficas y los códigos geográficos. El registro de los domicilios debe ser de buena calidad, de modo que los inmuebles residenciales y los establecimientos de las empresas puedan ligarse a domicilios estandarizados.

Los mapas digitales que definen la ubicación de todos estos objetos se pueden considerar registros y también se deben incluir en el registro base. Estos mapas se deben elaborar a partir de las coordenadas disponibles en el Catastro Nacional para así definir la dimensión geográfica en el sistema de registros. Hoy día estos mapas serán mapas con puntos de coordenadas. Si el mapa de polígonos de los bienes inmuebles está disponible, también se debe incluir.

Un registro base debe definir las poblaciones de objetos. Así pues, las variables importantes son aquellas que identifican y ubican los objetos, y que definen momentos de los eventos relevantes relacionados con esos objetos. Un registro de ese tipo se puede integrar con ayuda del registro de tierras y edificaciones del Catastro Nacional y complementarse con un registro de viviendas y un registro tributario de bienes inmuebles. Asimismo, un registro base inmobiliario debe incluir vínculos con las unidades tributarias y edificaciones. Entonces el registro base se puede actualizar con datos sobre catastro, permisos de edificación y tributación de bienes inmuebles.

2.2.5 ¿Por qué hay cuatro registros base?

Algunos países tienen una estructura compuesta por tres registros base: personas, empresas y bienes inmuebles. Esta estructura se refiere a los *registros administrativos* que se usan para administrar estos tres tipos de objetos. Por consiguiente, es correcto decir que hay tres registros *administrativos* importantes o básicos que se usan para el pago de impuestos de personas, empresas y bienes inmuebles. Sin embargo, debemos distinguir entre fines administrativos y fines estadísticos, pues una buena estructura administrativa no necesariamente será una buena estructura estadística.

Las actividades no son objetos administrativos —las verificaciones de ingresos, en las que se sustenta el segmento correspondiente a empleo del Registro de Actividades, se usan para la tributación de personas—. No obstante, las actividades son objetos estadísticos interesantes y existen muchas otras variables con relevancia estadística, como salario y ocupación, que describen estas actividades. Como una persona puede tener varios empleos y actividades de estudio durante un año calendario, el Registro de Actividades no se debe reducir a un registro sobre personas, porque se produciría una grave pérdida de información. Por este motivo las actividades se incluyen en un cuarto registro base dentro del sistema de registros *estadísticos*.

El modelo conceptual del sistema de registros estadísticos debe mostrar cómo "opera" el sistema. Las líneas del gráfico 2.5 indican los vínculos entre todos los registros del modelo. Estas líneas o vínculos muestran la manera en que los datos se pueden integrar al sistema. Para mostrar de qué manera los datos se integran al sistema, necesitamos un modelo con cuatro registros base y cuatro vínculos básicos, es decir, los vínculos entre los registros base.

Por consiguiente, el Registro de Actividades es necesario como vínculo entre personas y empresas. Un modelo de registro con cuatro registros base posee una estructura clara y destaca las *actividades* como un conjunto de objetos de interés estadístico que se puede ilustrar mediante diversas fuentes administrativas. El modelo se usa para dar una idea clara del sistema y mostrar cómo se articulan las distintas partes. Un buen modelo de registro *estadístico* debe entonces constar de cuatro, no tres, registros base.

2.3 EL SISTEMA DE REGISTROS EN SU CONJUNTO

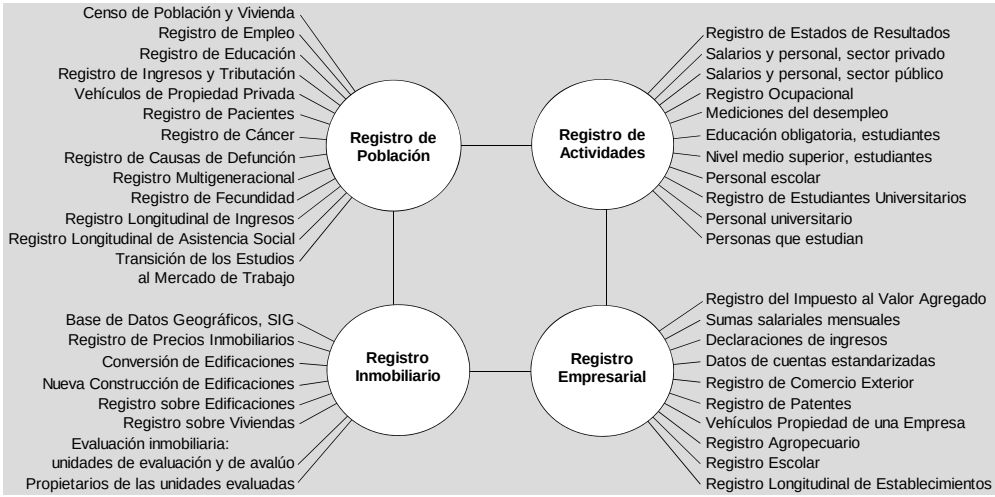
El siguiente paso consiste en vincular los diversos registros estadísticos con el registro base respectivo. En el gráfico 2.10 los círculos representan los registros base y las líneas, los vínculos entre objetos de diferentes registros.

Un sistema eficaz requiere que se hayan coordinado las definiciones de población y objetos, que se hayan armonizado las referencias temporales y las variables comunes, y que haya buenas variables de identificación, útiles para vincular objetos en los diferentes registros.

En el gráfico 2.10 vemos el contenido de un sistema de registros estadísticos bien diseñado que se basa en múltiples fuentes administrativas y algunos censos.

Algunos de los registros del sistema se generaron integrando información que ya existía en el sistema. Los registros longitudinales son ejemplos de estos registros integrados.

Gráfico 2.10 Sistema de registros estadísticos por tipo de objeto y campo de estudio



Nota: Éste no es un modelo del sistema de la Oficina de Estadística de Suecia, sino que muestra posibilidades generales. El sistema sueco contiene más registros, pero la Oficina de Estadística de Suecia no es responsable de todos ellos.

En la vida real, los objetos nacen, cambian y mueren todo el tiempo. Por ello, los registros del sistema cambian constantemente: se agregan y eliminan objetos. Las relaciones entre objetos se modifican y, en consecuencia, cambian las propiedades y los valores de las variables de los objetos. Cada uno de los registros incluidos en el modelo anterior existe para diferentes años y puede tener varias versiones.

En general, el sistema de registros se puede alimentar de manera diaria, semanal, mensual o anual. Una gran parte del sistema de registros sueco recibe datos mensualmente o incluso con mayor frecuencia, por lo que algunas estadísticas basadas en registros pueden ser muy oportunas.

2.4 CONSTRUCCIÓN Y USO DEL SISTEMA

La parte del sistema de registros relacionada con las estadísticas sobre personas es la más desarrollada en Suecia y es útil para mostrar las posibilidades que ofrece un sistema coordinado.

2.4.1 Estadísticas sobre personas basadas en el sistema de registros sueco

Los registros y productos que aparecen en el gráfico 2.11 ilustran paso a paso cómo crear un sistema y cómo darle varias aplicaciones. El primer paso consiste en idear aplicaciones a partir del registro base. Un registro base contiene objetos y conjuntos de objetos importantes, y las estadísticas demográficas al respecto constituyen una interesante aplicación.

El segundo paso es idear aplicaciones usando cada fuente administrativa, una aplicación por cada fuente principal. El siguiente paso es desarrollar aplicaciones más avanzadas integrando datos de diferentes registros estadísticos ya presentes en el sistema. Algunos de estos registros integrados pueden servir para estadísticas oficiales y otros para investigaciones académicas.

También hay algunas aplicaciones especiales que usan registros. Tanto en la investigación como en el gobierno se usan modelos de microsimulación y hay una gran demanda de tablas regionales detalladas.

Gráfico 2.11 Estadísticas sobre personas basadas en el sistema de registros sueco

1. Registro base:	*Estadísticas demográficas:* población, nacimientos/defunciones, migración *Encuestas por muestreo:* el Registro de Población se usa como marco muestral
2. Registros basados directamente en datos administrativos:	Registro de Ingresos y Tributación Vehículos de Propiedad Privada Registro de Pacientes Registro de Cáncer Registro de Causas de Defunción
3. Registros integrados para estadísticas oficiales:	"Censo" basado en registros Registro de Empleo Registro de Educación *Encuestas por muestreo:* En vez de preguntar, por ejemplo, la edad, el ingreso y el nivel de escolaridad, se usan variables incluidas en los registros del sistema (los registros mencionados en los incisos 1, 2 y 3). Muchas variables del sistema de registros se pueden usar como variables auxiliares para estratificación y calibración de los ponderadores de muestreo.
4. Registros integrados para investigación:	Registro Multigeneracional Registro de Fecundidad Registro Longitudinal de Ingresos Registro Longitudinal de Asistencia Social Transición de los Estudios al Mercado Laboral
5. Modelos de microsimulación:	El Registro de Ingresos y Tributación constituye la base de un modelo de simulación donde se pueden poner a prueba los cambios planeados en tributación y pagos por concepto de transferencias. Este modelo se usa tanto en el gobierno como en la investigación académica. El Registro de Empleo constituye la base de un modelo regional de simulación y pronóstico usado en el gobierno y la investigación académica.
6. Tablas regionales estandarizadas:	El Registro de Población, el Registro de Ingresos y Tributación, los Vehículos de Propiedad Privada, el Registro de Empleo y el Registro de Educación conforman la base de las tablas regionales estandarizadas que usan los gobiernos locales y los analistas del mercado. Las regiones se pueden seleccionar mediante tecnología SIG.

2.4.2 Consistencia respecto a las poblaciones y variables

En el gráfico 1.1 presentamos cuatro principios sobre el uso de los datos administrativos:

1. Una oficina de estadística debe tener acceso a los registros administrativos que mantengan las autoridades. Este derecho, al igual que la protección de la privacidad, debe estar amparado por ley.
2. Los registros administrativos se deben transformar en registros estadísticos. Durante este proceso, se usarán y compararán fuentes muy diversas.
3. Todos los registros administrativos se deben incluir en un sistema de registros coordinado. Este sistema asegurará que todos los datos se puedan integrar y usar eficazmente.
4. La consistencia en las poblaciones y las variables es necesaria para que las estimaciones de diferentes encuestas basadas en registros sean coherentes.

En las secciones 1.4.1, 1.4.2 y 1.5.4 ejemplificamos el segundo principio y en este capítulo ya ilustramos anteriormente el tercero. La Oficina de Estadística de Dinamarca nos inspiró esos dos principios. Ahora explicaremos el cuarto principio, inspirado por la Oficina de Estadística de los Países Bajos.

La consistencia respecto a las poblaciones se logra si todas las encuestas basadas en registros usan la misma versión del registro base como población. La unidad responsable del Registro de Población crea una versión especial cada año definida como el registro que describe la población al 31 de diciembre. Este registro se usa como una población estandarizada y define el conjunto de objetos para muchos otros registros con datos sobre personas. Este procedimiento asegura la consistencia entre los diferentes productos de registros en lo concerniente a las poblaciones.

La responsabilidad de las variables importantes de interés común se descentraliza de modo que las personas que trabajan con los diferentes registros asuman la responsabilidad de *sus* variables. El conocimiento detallado de los temas específicos es de suma importancia en este trabajo, tanto para el análisis de las fuentes administrativas con las autoridades que entregan los datos como para el procesamiento de las variables administrativas y la creación de las variables estadísticas. Ése es el principal motivo por el que se justifica una organización descentralizada. Como vemos en el siguiente gráfico, cuatro distintas unidades de registro de la Oficina de Estadística de Suecia son responsables de *sus* variables estandarizadas. Las variables estandarizadas, concepto que abordaremos en la siguiente sección, son variables importantes para el sistema de registros como un todo.

Gráfico 2.12 Variables estandarizadas: datos sobre personas

Unidad de registro responsable del:	Responsabilidad, variables estandarizadas
Registro de Población	Variables demográficas, estado civil, país de nacimiento
Registro de Empleo	Situación de empleo
Registro de Educación	Nivel de escolaridad final
Registro de Ingresos y Tributación	Diferentes variables sobre los ingresos

La responsabilidad de las variables está descentralizada, pero las unidades de registro cooperan y usan la misma población. Esto imprime consistencia a las estadísticas basadas en registros, como vemos en el siguiente gráfico, que contiene datos de los cuatro registros antes mencionados.

Gráfico 2.13 Estadísticas basadas en registros para un municipio pequeño, 2003

Población		**Empleo**		**Nivel de escolaridad**					**Ingreso** Anual percibido, en miles ($)				
Edad	Número	Empleado	Desem-pleado	Básico	Medio superior	Superior	Posgrado	No especifi-cado	0	1–14	15–29	30–44	45+
0–15	1416	-	-	-	-	-	-	-	-	-	-	-	-
16–19	387	69	318	306	71	0	0	10	118	265	4	0	0
20–24	293	207	86	44	219	26	0	4	12	130	128	23	0
25–34	764	616	148	79	469	210	0	6	20	133	388	202	21
35–44	937	782	155	142	558	226	2	9	27	128	440	270	72
45–54	1002	847	155	259	510	225	4	4	14	90	501	318	79
55–64	1042	713	329	420	413	199	6	4	21	166	502	288	65
65+	1199	40	1,159	333	168	78	3	617	3	552	535	90	19

Los cuatro registros del gráfico 2.12 también podrían considerarse un gran registro con todas las variables, creado mediante un proceso descentralizado, pero coordinado. En el gráfico 2.13, se describe cada una de las variables. Sin embargo, a partir del gran registro integrado con todas las variables podrían elaborarse muchos tabulados y análisis multidimensionales. El registro integrado sería completamente consistente en relación con las variables, es decir, no habrá información contradictoria porque los cuatro registros son idénticos.

Los usuarios agradecen la consistencia de las estadísticas generadas a partir de estos registros. Hay una total consistencia en estas estadísticas, que describen alrededor de 300 municipios, 100 grupos de edad y dos géneros, lo que equivale a una tabla de 60 000 celdas. Una condición que debe cumplirse para producir este tipo de estadísticas coordinadas a partir de registros es que el registro base, en este caso el Registro de Población, sea de calidad alta.

El registro integrado de nuestro ejemplo viene de la Social Statistics Database (SSD), estudiada por varios autores de la Oficina de Estadística de los Países Bajos (2000) y por Laan (2002). Estos autores abordan el emparejamiento exacto y la microintegración de los datos de muchas encuestas estadísticas, ya sea encuestas basadas en registros, censos o encuestas por muestreo. Tienen la visión de una consistencia total en cuanto a la población y las variables. Los registros suecos que mencionamos antes también son consistentes en estos aspectos y pueden integrarse en el nivel micro mediante un emparejamiento exacto. En el capítulo 13 abordaremos más a fondo los conceptos de coordinación, coherencia y consistencia.

Para crear un sistema de registros coordinado en el que se puedan integrar todos los registros es necesario que los registros base estén coordinados. Si se crean versiones por año calendario de los registros base, entonces:

– las personas vinculadas con las actividades del Registro de Actividades también deben existir en el Registro de Población;

– las organizaciones y establecimientos incluidos en el Registro de Actividades también deben existir en el Registro Empresarial;

– los domicilios de los establecimientos incluidos en el Registro Empresarial deben hallarse en el Registro Inmobiliario;

– las viviendas y los inmuebles donde habitan las personas incluidas en el Registro de Población deben encontrarse en el Registro Inmobiliario.

2.5 VARIABLES ESTANDARIZADAS EN EL SISTEMA DE REGISTROS

Ciertas variables se usan en múltiples registros, por lo que son fundamentales para todo el sistema. Por ejemplo, una *variable estandarizada*[3] reviste tal importancia que se toma una decisión especial respecto a la responsabilidad correspondiente. Los encargados de un registro del sistema que reciben esa variable de una autoridad administrativa o que la crean deben hacerse totalmente responsables de ella en la oficina de estadística. Tener plena responsabilidad por una *variable estandarizada* implica:

– mantenerse en contacto con la autoridad que suministra el registro administrativo y asegurarse de que la oficina de estadística recibe de esa autoridad tanto la variable como la documentación sobre ella;

– ocuparse de su validación, procesamiento, designación y documentación.

Dentro de la oficina de estadística, los encargados de otros registros que usan una variable estandarizada deben utilizar dicha variable sin modificaciones, con el nombre correcto y la documentación original. De este modo, se minimizará la duplicación del trabajo, se facilitará la coordinación y mejorará la consistencia al interior del sistema. Hay dos categorías de variables estandarizadas que tienen un papel especial en el sistema: los *vínculos* y las *clasificaciones.*

2.5.1 Vínculos

En el gráfico 2.5 se muestran los vínculos que constituyen el fundamento del sistema, basados en las siguientes identidades: *número de identificación personal, número de identificación de organización o unidad jurídica, número de identificación de establecimiento,* e *identidad inmobiliaria y habitacional.* Además, el *domicilio de ubicación de un establecimiento* es el vínculo entre el Registro Empresarial y el Registro Inmobiliario.

Cuando se crea un nuevo registro en el sistema, se lleva a cabo un emparejamiento exacto entre los distintos registros con ayuda de variables identificadoras. Un registro administrativo se empareja con un registro base para verificar los conjuntos de objetos. Además, se efectúan otros emparejamientos con distintos registros estadísticos a fin de reunir variables para el nuevo registro. En todo este proceso de emparejamiento se da por sentado que todos los registros contienen vínculos apropiados. Un vínculo consiste en una o más variables que identifican objetos independientes. Es importante que la misma variable identificadora esté presente en muchos registros, para facilitar el emparejamiento. Cuando se empareja un registro que se relaciona con diferentes años, son necesarios vínculos que se mantengan estables a lo largo del tiempo, es decir, un objeto específico que no cambia debe tener los mismos valores para las variables identificadoras.

Ejemplo: El número de identificación personal se introdujo en Suecia en 1947. Es un número de identidad nacional creado por el Consejo Nacional Tributario para todas las personas que residen de manera permanente en Suecia y registradas ante esa oficina; se usa en casi todos los sistemas administrativos. Se trata de una variable muy estable de gran calidad, lo que facilita el procesamiento del registro.

Ejemplo: En Suecia, la identidad inmobiliaria está formada por el código del condado, del municipio y del inmueble. Éste es un ejemplo de una variable muy inestable. Si el código del municipio o el condado cambian debido a modificaciones en las divisiones

[3] En los libros de texto se usa el mismo término para otro concepto: una variable estandarizada tiene una media de 0 y una desviación estándar de 1.

administrativas, también cambiarán muchas identidades inmobiliarias. Sería mucho mejor usar números de identidad, que serán los mismos siempre que los inmuebles correspondientes no sufran modificaciones.

2.5.2 Clasificaciones

La rama industrial, la categoría de producto, el nivel de escolaridad y la ocupación son algunos ejemplos de *estándares* y *clasificaciones* estadísticos importantes, basados en recomendaciones internacionales. Su importancia radica en el contenido que permite que sean usados en muchas encuestas, tanto basadas en registros como de otro tipo. En muchos casos, las fuentes administrativas contienen datos sobre estas clasificaciones, información que se aprovecha para crear variables dentro del sistema de registros. Además, los responsables de un *estándar* en una oficina de estadística deben estar habilitados para *codificar* nuevos términos referidos a ocupaciones, nuevos programas educativos, etcétera, de tal modo que los nuevos términos se incluyan en una categoría adecuada dentro de las clasificaciones.

2.6 SISTEMAS DE REGISTROS ESTADÍSTICOS FUERA DE LA OFICINA DE ESTADÍSTICA DE SUECIA

En muchos países, las autoridades en materia de estadística muestran un interés cada vez mayor en el aprovechamiento de fuentes administrativas para sus estadísticas. Las estadísticas internas dentro de las grandes empresas también se basan completamente en registros.

2.6.1 Estadísticas basadas en registros en otros países

Los países escandinavos poseen una larga tradición de estadísticas basadas en registros. Cuando Dinamarca, Finlandia y Suecia se integraron a la Unión Europea (UE), dio inicio el debate sobre este "nuevo" método para la producción de estadísticas oficiales dentro de Eurostat, organismo responsable de la cooperación estadística en la UE.

En las actas de un seminario organizado por este organismo en 1997 (Eurostat, 1997), se describe la situación prevaleciente. El seminario representó un punto de inflexión, pues a partir de entonces las estadísticas basadas en registros ganaron aceptación como método para producir estadísticas oficiales dentro de la UE. En la sección 2.4.2 mencionamos a la Oficina de Estadística de los Países Bajos, donde están elaborando bases de datos sociales y económicos, en buena medida aprovechando la información de fuentes administrativas.

En sus Directrices de Calidad, la Oficina de Estadística de Canadá (2003) establece una distinción entre varios tipos de encuestas: censos, encuestas por muestreo y captación de datos incluidos en registros administrativos, es decir, lo que nosotros denominamos encuestas basadas en registros. La política de la Oficina de Estadística de Canadá es aprovechar los registros administrativos siempre que ello ofrezca una mejor relación costo-beneficio en comparación con la captación de datos directa. Se usan cada vez más fuentes administrativas, sobre todo para las encuestas empresariales. En la Oficina de Estadística de Canadá (2006), se describe el aprovechamiento actual de los datos administrativos para empresas. En ese artículo se analiza un planteamiento sobre sistemas que tiene por objeto producir estadísticas económicas consistentes con base en el Registro Empresarial, lo que se asemeja a la función que consideramos deben tener los registros base.

Sin embargo, los datos administrativos de la Agencia Tributaria de Canadá no se usan para crear registros, sino como sustitutos de los datos de encuestas por muestreo para pequeñas empresas y para una fracción de la muestra a fin de reducir la carga de respuestas.

De acuerdo con el Plan Estratégico de la Oficina del Censo de Estados Unidos (2003), la carga para los entrevistados y el costo para los contribuyentes se deben minimizar obteniendo y elaborando datos de alta calidad a partir de fuentes que mantienen otras entidades gubernamentales y comerciales. Otro de los objetivos consiste en producir nueva información aprovechando las fuentes de datos existentes mediante una microintegración. "Una de las grandes oportunidades para mejorar las mediciones estadísticas actuales está en integrar varias fuentes de información". Si deseamos alcanzar esa meta, hace falta investigar cómo mejorar los métodos.

En el Servicio de Impuestos Internos (IRS, por sus siglas en inglés) de Estados Unidos, las declaraciones de personas y empresas se usan para propósitos estadísticos. Scheuren y Petska (1993) ofrecen un panorama de las operaciones estadísticas en el IRS. Esas operaciones se sustentan en una tradición larga y bien establecida de cómo aprovechar las fuentes administrativas, por lo que haremos algunas comparaciones respecto a nuestras experiencias en Suecia.

En el modelo del sistema de registros que aparece en el gráfico 2.10 hay tres registros importantes que corresponden a los que mantiene el IRS:

– El Registro de Ingresos y Tributación sobre personas y unidades familiares tributarias se basa en las declaraciones de impuestos de personas. Este tipo de registro también se usa para los modelos de microsimulación tanto del IRS como de la Oficina de Estadística de Suecia.
– El Registro de Estados de Resultados corresponde a los "documentos informativos" mencionados por Scheuren y Petska. Los proveedores de diferentes tipos de ingresos (empleadores, bancos, etc.) notifican los ingresos de cada persona con un empleo remunerado. Los informes de los empleadores son muy importantes, pues los datos que proporcionan son el vínculo entre las personas y empresas en el sistema de registros. Como señalamos en la sección 2.2.2, esta información puede usarse para crear el Registro de Empleo, que constituye la mayor parte del Registro de Actividades.
– Datos de cuentas estandarizadas. Este registro se basa en las declaraciones de impuestos de diferentes tipos de empresas.

Nuestra conclusión es que existen los mismos tipos de fuentes administrativas importantes en Estados Unidos y los países escandinavos, y que estas fuentes se han aprovechado para fines estadísticos desde hace mucho tiempo. En Estados Unidos, por sus dimensiones, sólo se usa una muestra de todas las declaraciones de impuestos para el análisis estadístico del IRS. Las declaraciones que integran la muestra se validan y los valores perdidos se remplazan por valores imputados. En Suecia, se hace un muestreo de los registros administrativos para el modelo de microsimulación, que se describe en la sección 6.3.1. Pensamos que con el avance de la informática, el muestreo se volverá innecesario y la práctica común será la de los registros de ingresos estadísticos basados en las declaraciones de impuestos.

Scheuren (1999) analiza el uso de registros administrativos en aplicaciones de censos. Compara las condiciones de Canadá, Estados Unidos y Europa. Estados Unidos enfrenta el mayor desafío. Desde luego, existen muchas fuentes administrativas en ese país, de modo que un censo se puede basar parcialmente en registros. No obstante, debe haber leyes que obliguen a la cooperación de distintas dependencias, como la Ley de Estadística de Canadá.

Asimismo, los ciudadanos deben acostumbrarse al remplazo de las entrevistas y cuestionarios por la vinculación de registros administrativos.

En una serie de sesiones de trabajo organizadas por la Comisión Económica de las Naciones Unidas para Europa (UNECE, por sus siglas en inglés) y Eurostat, se analizaron los censos de población y vivienda basados en registros. En UNECE (1998) se presentan las recomendaciones para los censos de la ronda 2000. El método de aprovechamiento de registros y otras fuentes administrativas, de ser necesario en combinación con encuestas por muestreo, está aceptado por la UNECE.

El Censo de Población y Vivienda de 1981 de Dinamarca fue el primer censo del mundo que se llevó a cabo total y exclusivamente con base en información de registros administrativos. Desde hace largo tiempo, la Oficina de Estadística de Dinamarca considera que las encuestas basadas en registros son la metodología estadística más importante y con la mejor relación costo-beneficio. El primer libro sobre encuestas basadas en registros se publicó en 1994 (en danés), al que siguió una versión en inglés (Statistics Denmark, 1995).

La Oficina de Estadística de Finlandia (Statistics Finland, 2004) hizo una descripción de las estadísticas basadas en registros en ese país. Desde 1990, esta oficina logró realizar censos de población y vivienda basados completamente en registros.

En el gráfico siguiente, indicamos qué partes del sistema de registros mostrado en el gráfico 2.10 se usan en un censo de población y vivienda basado en registros.

Gráfico 2.14 Registros estadísticos usados en un censo basado en registros

Registros censales	Registro usados en la generación de registros censales
Registro de Población	
Registro de Empleo	Registro de Estados de Resultados (registro de empleo basado en pagos de impuestos anticipados) Registro Empresarial
Registro de Educación	Escolaridad obligatoria Nivel medio superior Registro de Estudiantes Universitarios
Registro de Ingresos y Tributación	
Registro Inmobiliario	
Registro de Vivienda	

Como mencionamos antes, la Oficina de Estadística de los Países Bajos integró la Social Statistics Database (SSD). Se usan muchas fuentes administrativas para crear registros estadísticos, lo que corresponde a los registros antes mencionados. El censo virtual neerlandés de 2001 se realizó a partir de la SSD, a su vez basada en una combinación de registros y encuestas por muestreo.

Las fuentes aprovechadas y un esquema de la metodología aplicada para ese censo virtual se describen en Statistics Netherlands (2004). En el capítulo 4, explicamos la combinación de registros y encuestas por muestreo.

Bethlehem *et al*. (2006) ofrecen una descripción de la estrategia neerlandesa para pasar de la estadística basada en encuestas por muestreo a la basada en registros, así como del modelo del sistema de registros neerlandés. Señalan que las necesidades de los clientes están cambiando y ha aumentado la demanda de publicaciones temáticas en donde se combinen datos de varias fuentes. También hay una demanda creciente de estadísticas

regionales detalladas. Las encuestas basadas en registros donde se integran datos de fuentes múltiples pueden satisfacer estas demandas.

Fuentes administrativas importantes

Las oficinas de estadística de otros países a menudo trabajan con encuestas referidas a los mismos tipos de objetos que el sistema de registros de la Oficina de Estadística de Suecia. Esto significa que sería posible estructurar los sistemas de registros disponibles o planeados de la misma manera que el sistema sueco, como se muestra en las gráficas 2.5 y 2.10. Las principales fuentes administrativas que sustentan un sistema de este tipo aparecen en el gráfico 2.4:

– impuesto sobre la renta de personas, que incluyen las personas registradas en un domicilio, inmueble o vivienda;
– impuesto predial de tierras e inmuebles;
– impuesto sobre la renta e impuesto sobre bienes y servicios (o impuesto al valor agregado) de empresas;
– fuentes administrativas que vinculan a empleadores y empleados.

En muchos países, las estadísticas producidas se estructuran tradicionalmente en dos subsistemas, uno para estadísticas personales o sociales y otro para estadísticas económicas. Por consiguiente, a menudo se soslayan las posibilidades de integrar información de registros sobre personas y registros sobre empresas.

Con un sistema de registros como el del gráfico 2.10, es posible describir a *personas* mediante las características de las *empresas* para las que trabajan: personas empleadas por rama industrial, por tamaño de la empresa, por edad de la empresa, etcétera.

Por su parte, la *empresa* se puede describir mediante las características de las *personas* que trabajan para ella: proporción de hombres y mujeres, proporción de jóvenes, proporción de personas con un alto nivel de escolaridad, etcétera.

Para satisfacer la necesidad de estadísticas de este tipo, los registros sobre personas y los registros sobre empresas se deben incorporar en *un solo* sistema coordinado.

Números de identificación nacionales y otras identidades

Las oficinas de estadística de los países escandinavos tienen la ventaja de que desde hace muchos años un gran número de sistemas administrativos nacionales usan números de identificación nacionales únicos. Todos los sistemas utilizan los mismos NIP únicos para personas y los mismos números de identificación empresarial (NIE) únicos para empresas (unidades jurídicas). También hay números de código domiciliar y números de identificación inmobiliaria únicos. Todos estos números de identificación nacionales facilitan el uso eficiente de datos administrativos para propósitos estadísticos.

Sin embargo, estos números de identificación no son tan perfectos como pensarían los estadísticos fuera de los países escandinavos: dos objetos pueden tener el mismo número de identificación, el mismo objeto puede tener dos números de identificación, etc. Esto significa que siempre habrá, hasta cierto punto, errores y problemas de emparejamiento. Los problemas de emparejamiento suelen ser engorrosos cuando se ha creado un nuevo registro o se usan fuentes administrativas por primera vez. Estos problemas se atenuarán con el tiempo —el mejoramiento de la calidad es una tarea de largo plazo que poco a poco reducirá los problemas de emparejamiento—.

Ejemplo: En la primera versión del Registro de Empleo sueco (1985) fue posible vincular 93.6% de las personas empleadas con el establecimiento correspondiente en el Registro Empresarial. Nueve años después, este porcentaje se elevó a 98.3 por ciento.

En los países donde no hay números de identidad nacionales, deben usarse otros tipos de identidad en el sistema de registros estadísticos. Datos como nombre, domicilio, fecha de nacimiento, lugar de nacimiento, etc. son posibles variables de vinculación. Las oficinas nacionales de estadística deben dar seguimiento a los cambios de nombre y domicilio para usar esas variables como variables de vinculación. Asimismo, pueden crear sus números de identidad propios con base en las variables de vinculación disponibles. Estos números se usan después como vínculos en las versiones finales de todos los registros estadísticos pertenecientes al sistema de registros.

El emparejamiento dentro del sistema se lleva a cabo con los vínculos del sistema. El propósito del proceso de emparejamiento consiste en hallar objetos idénticos en diferentes registros o en hallar objetos con una relación definida. Estos vínculos se crean con las variables de identificación disponibles.

Como cada país puede tener sus propias variables de identificación, distintas de las de otros países, cuando se crean estos vínculos, los problemas de emparejamiento diferirán en cierta medida, pero se puede aplicar la misma teoría a los sistemas de registros estadísticos y las encuestas basadas en registros.

2.6.2 Estadísticas basadas en registros dentro de las empresas

Las empresas y organizaciones cuentan con sus propios sistemas de registros administrativos, que también pueden aprovecharse para propósitos estadísticos. Las empresas manufactureras tienen sistemas administrativos formados por dos partes: una para finanzas y otra para gestión de materiales y producción. Estos sistemas contienen cientos de registros con miles de variables.

Así como una oficina de estadística utiliza registros administrativos para crear registros estadísticos, las empresas pueden aprovechar sus registros administrativos para generar registros que sirvan como fuentes para estadísticas internas, por ejemplo, sobre ventas. Estos registros estadísticos contienen microdatos para todas las operaciones relacionadas con los nuevos pedidos y la facturación.

En el sistema de registros de una empresa manufacturera, ciertos registros se pueden considerar registros base. El registro de artículos y el de clientes son dos ejemplos que definen tipos de objetos importantes, de modo que la identidad del cliente y la del artículo serían dos variables de vinculación relevantes en el sistema.

Así como los sistemas administrativos de la sociedad pueden usarse para fines estadísticos, también sería posible investigar el aprovechamiento estadístico de los datos administrativos de una empresa. Las ciencias estadísticas deben contribuir al desarrollo en este campo. El rápido crecimiento de la *minería de datos* refleja esta necesidad.

Para ilustrar lo anterior recurriremos al ejemplo de una encuesta basada en registros en una empresa manufacturera. Se trata de la encuesta mensual sobre ventas. Cada mes, se crea un registro estadístico mediante emparejamientos y selecciones de tres registros administrativos: el registro de facturación, que incluye todas las operaciones relacionadas con facturas, el registro de clientes y el registro de artículos.

Gráfico 2.15 Tres registros administrativos

Registro de facturación

Fecha	Núm. de cliente	Núm. de artículo	Cantidad	Valor
2006-01-18	196	22	10	832
2006-01-19	28	4	500	20 339
2006-01-19	7	128	40	9 840
2006-01-20	23	9	100	10 622

Registro de clientes

Núm. de cliente	Segmento	País
7	3	SUE
23	3	SUE
28	3	SUE
196	2	GB

Registro de artículos

Núm. de artículo	Grupo de artículos	Costo previsto
4	1	36
9	1	90
22	2	28
128	2	205

Se crea un registro estadístico de ventas de la siguiente manera:

Del registro de facturación se seleccionan como objetos del registro de ventas todas las operaciones para determinado mes. Ese registro mensual se empareja con los registros de clientes y artículos y las variables de esos registros se importan al registro de ventas.

Gráfico 2.16 Registro de ventas para enero de 2006: cuatro operaciones

Fecha	Núm. de cliente	Segmento	País	Núm. de artículo	Grupo de artículos	Cantidad	Valor	Precio	Costo previsto	Utilidades brutas
2006-01-18	196	2	GB	22	2	10	832	8.32	280	552
2006-01-19	28	3	SUE	4	1	500	20 339	40.68	18 000	2 339
2006-01-19	7	3	SUE	128	2	40	9 840	246.00	8 200	1 640
2006-01-20	23	3	SUE	9	1	100	10 622	106.22	9 000	1 622

El registro de ventas para determinado mes se usa entonces para crear tablas con los valores facturados a los precios actuales, índices de precios, volúmenes facturados (valores a precios constantes) y márgenes de utilidades brutas por segmento, país y grupo de artículos. Estas tablas se usan para actualizar una base de datos de series de tiempo.

Los registros administrativos deben transformarse en registros estadísticos. En este caso, los datos pueden ser de buena calidad administrativa, pero no adecuados para fines estadísticos.

Los dos ejemplos siguientes se refieren a la importancia de transformar los datos administrativos, pues los errores en esos datos se deben corregir de acuerdo con principios estadísticos y los valores perdidos se deben tratar como valores perdidos en datos estadísticos.

1. Correcciones en datos contables

El 27 de febrero se consigna una facturación en el registro de facturación. Cuando se verifican las operaciones de febrero se observa que esta operación está equivocada y no debió realizarse. Se hace una corrección el 2 de marzo.

Gráfico 2.17

Registro administrativo de facturación

Fecha	Núm. de cliente	Núm. de artículo	Cantidad	Valor
2006-02-27	53	9	1 000	107 560
2006-02-28	34	112	655	32 700
2006-03-01	117	4	500	20 339
2006-03-02	53	9	–1 000	–107 560

Registro estadístico de facturación

Fecha	Núm. de cliente	Núm. de artículo	Cantidad	Valor
2006-02-28	34	112	655	32 700
2006-03-01	117	4	500	20 339

De acuerdo con los principios de contabilidad, los errores se corrigen agregando una nueva operación de modo que la operación errónea y la corrección sumen cero. Desde el punto de vista estadístico, esto corresponde a dos errores con diferente signo. En el error estadístico

ni el error ni la corrección deben incluirse porque entonces la serie de tiempo será de baja calidad estadística cuando el error y la corrección aparezcan en meses distintos.

2. Valores perdidos

Los registros administrativos pueden contener valores perdidos. Si éstos se interpretan como ceros, los análisis estadísticos serán engañosos. Cuando se genera el registro estadístico deben detectarse los valores perdidos y sustituirse por valores imputados.

Gráfico 2.18 Registro administrativo de artículos

Núm. de artículo	Grupo de artículos	Costo previsto
4	1	36
9	1	
22	2	28
128	2	205

Puede ocurrir que no se hayan calculado los costos previstos para meses anteriores.

Si no se detectan estos valores perdidos, la tendencia de los márgenes brutos será equivocada.

Los datos administrativos deben validarse antes de su uso para fines estadísticos.

Gráfico 2.19 Registro estadístico de artículos

Núm. de artículo	Grupo de artículos	Costo previsto
4	1	36
9	1	94.42
22	2	28
128	2	205

Las ventas totales de artículos con costos previstos conocidos durante el periodo ascienden a 90 000 coronas suecas.

Los costos previstos para estas ventas ascienden a 80 000 coronas.

El precio promedio del artículo 9 durante el periodo es de 106.22 coronas.

Por consiguiente, el costo previsto imputado para el artículo 9 es:
106.22 · (80 000/90 000) = 94.42

Las encuestas basadas en registros son comunes dentro de empresas y otras organizaciones. Así como en el resto del libro veremos la manera en que las oficinas de estadística pueden aprovechar datos administrativos para producir estadísticas oficiales, los problemas metodológicos asociados con estas encuestas empresariales deben investigarse y analizarse. Con frecuencia estos problemas no se reconocen como problemas estadísticos porque las ciencias estadísticas sólo se asocian con las encuestas por muestreo, los experimentos aleatorizados, la probabilidad y la teoría de la inferencia.

CAPÍTULO 3

Una terminología para encuestas basadas en registros

El desarrollo de las estadísticas basadas en registros requiere un vasto idioma común. Sabemos que hay un idioma propio de la teoría de las encuestas por muestreo. Términos como *marco muestral*, *muestreo probabilístico*, *estimador* y *error estándar* son muy conocidos y poseen una definición clara. Las estadísticas basadas en registros tienen la misma necesidad de una terminología bien establecida. Pensamos con palabras y si nuestros términos para la estadística basada en registros son poco claros, lo mismo ocurrirá con nuestro pensamiento. En cambio, si tenemos una serie de conceptos adecuados y bien definidos, se fomentará el intercambio de conocimientos y la formulación de metodologías.

Todos los términos que proponemos se compilan en el glosario incluido al final del libro.

3.1 TERMINOLOGÍA: UN IDIOMA DIFERENTE

El procesamiento de registros ha recibido la influencia de la teoría de la informática, por lo que a menudo se usan términos de ese campo en vez de términos estadísticos. En ocasiones, el significado de los términos puede diferir dependiendo de que se usen en el contexto de la informática o de la estadística. El objetivo de la terminología que presentamos en este libro es denominar conceptos importantes e introducir términos unívocos de modo que sea posible formular una teoría común de los registros.

3.1.1 Conceptos y términos

Un *concepto* es el contenido abstracto de un *signo* lingüístico. Puede haber varios términos referidos al mismo concepto. Más adelante mostramos algunos ejemplos de que los términos usados en la estadística y en la informática pueden diferir aunque se refieran al mismo concepto.

El rápido desarrollo de la informática ha hecho que ciertos términos de ese campo evolucionen junto con la tecnología. Esto puede ser motivo de confusión y tal vez no se advierta que se hace referencia al mismo concepto estadístico. Cuando la Oficina de Estadística de Suecia cambió las computadoras centrales por servidores de bases de datos, términos antiguos como *archivo plano* con *registros* y *posiciones* fueron sustituidos por *tabla de base de datos* con *renglones* y *columnas*. Es necesario establecer términos propios de la estadística basada en registros que sobrevivan a ese tipo de cambios técnicos. Por ello usamos términos como *matriz de datos* en vez de *hoja de cálculo*, *conjunto de datos*, *tabla de base de datos* o *archivo plano*.

3.1.2 Encuesta estadística

En la sección 1.2 se mencionan diferentes tipos de encuestas: encuestas por muestreo, censos y encuestas basadas en registros, que son el fundamento para la producción de datos estadísticos en una oficina de estadística. Sin embargo, el término *encuesta* a menudo se usa para designar una encuestas por muestreo. Un ejemplo de ello es la definición para *encuesta* de la *American Statistical Association*[1], que claramente se refiere a una encuesta por muestreo. Asimismo, hay varios libros (confiables) donde algunos términos como *error de la encuesta*, *metodología de encuesta* y *calidad de la encuesta* sólo se refieren a las encuestas por muestreo.

¿Por qué debemos llamar *encuestas para propósitos estadísticos* a las *encuestas basadas en registros*? Porque su teoría también se debe basar en la metodología general para encuestas con propósitos estadísticos. En el caso de las encuestas basadas en registros, es necesario además definir los objetivos de investigación, la población y las variables, recopilar los datos (de una autoridad o del sistema de registros de la oficina de estadística) y trabajar con el aseguramiento de la calidad, el análisis y la presentación. Al usar el término *encuesta*, se ponen de relieve las semejanzas.

3.2 TERMINOLOGÍA DE LOS REGISTROS

Registro y *tabla* son términos importantes que pueden tener diferentes significados y no debe haber confusiones en su uso. En la sección 1.3 definimos de la siguiente manera el término *registro*: un registro tiene la finalidad de ser una *lista completa* de los objetos que forman parte de un grupo de objetos o una población específicos y debe contener información sobre la *identidad* del objeto, de modo que el registro se pueda actualizar con nuevos valores de las variable para ese objeto. Esta definición se aplica tanto a los registros administrativos como a los estadísticos.

3.2.1 Registro, matriz de datos y tabla

El ejemplo siguiente, con datos de un registro estadístico imaginario, nos servirá para explicar los términos *registro*, *matriz de datos* y *tabla.* Imaginemos que tenemos un registro que contiene datos sobre todas las empresas del país en determinado momento. Entonces el número de objetos en el registro, ilustrados en el gráfico siguiente, está dado por *N*, además el registro contiene seis variables.

Gráfico 3.1 Ejemplo de un registro y una matriz de datos

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*	*Variable 5*	*Variable 6*
	Nombre de la empresa	**Domicilio**	**Número de organización***	**Facturación**	**Empleados**	**Rama industrial**
Objeto 1	Pintores A	Domicilio 1	NIE 1	12	9	F
Objeto 2	Panadería B	Domicilio 2	NIE 2	3	4	D
Objeto 3	Peluquería C	Domicilio 3	NIE 3	7	10	O
.	.	.	.	.	.	.
Objeto N	Fábrica Z	Domicilio N	NIE N	211	76	D

* NIE: Número de Identificación de la Empresa.

[1] "A survey gathers information from a sample", *American Statistical Association* (1996).

En este sencillo ejemplo, hay una correspondencia entre la matriz de datos y el registro, pero normalmente no ocurre así. Compararemos estos conceptos en la sección 3.2.2. En una *matriz de datos*, los datos estadísticos se clasifican de modo que las columnas sean las variables y los renglones, las observaciones para los objetos. El registro del gráfico 3.1 está representado por una matriz de datos con *N* renglones y seis columnas. La matriz de datos se destaca con un recuadro y también agregamos un texto explicativo en la primera columna y el encabezado.

Toda encuesta para fines estadísticos tiene por objeto crear una o varias matrices de datos que contengan *microdatos*, que luego se procesarán para esos fines. El término *matriz de datos* se puede considerar un concepto estadístico para ese conjunto de datos.

Las columnas de la matriz contienen *mediciones*[2] de *variables*[3] y los renglones, *observaciones*[4] para los objetos del registro. La observación para el objeto 2 se marcó como renglón en blanco en el gráfico. En este caso, las observaciones son de seis dimensiones. Si obtenemos datos nuevos que incluyan, por ejemplo, un número de organización y una facturación revisada, el registro del gráfico 3.1 se puede actualizar formando una nueva versión en la que los antiguos valores se sustituyen por los nuevos.

Una matriz de datos contiene observaciones para objetos *individuales*, que reciben el nombre de microdatos. Si el registro es anonimizado, obtendremos la matriz de datos del gráfico 3.2. Esta matriz de datos no incluye ningún dato sobre la identidad de los objetos y no es posible agregar nuevas mediciones. Por consiguiente, *ya no* es un registro. En ocasiones se usa la expresión *registro anonimizado*, pero en realidad no es apropiado porque se contrapone a la definición del término *registro*.

Gráfico 3.2 Matriz de datos anonimizados

	Variable 1	*Variable 2*	*Variable 3*
	Facturación	**Empleados**	**Rama industrial**
Objeto 1	12	9	F
Objeto 2	3	4	D
.	.	.	.
Objeto N	211	76	D

Una muestra del registro del gráfico 3.1 nos da una tercera matriz de datos; sin embargo, como la muestra no es una lista completa de la población, esa matriz de datos no es un registro en el sentido que nosotros damos al término. En las matrices de datos generados con muestras, se puede considerar que cada observación en la muestra representa muchas observaciones en la población. En el gráfico 3.3, la primera observación representa 411.7 elementos en la población. Una parte importante del trabajo con las encuestas por muestreo consiste en calcular estos *ponderadores*.

Gráfico 3.3 Matriz de datos de una encuesta por muestreo

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*	*Variable 5*	*Variable 6*	*Variable 7*
	Nombre de la empresa	**Domicilio**	**Número de organización**	**Facturación**	**Empleados**	**Rama industrial**	**Ponderador**
Objeto 1	X S.A.	Domicilio 1	NIE 1	111	57	D	411.7
Objeto 2	Y compañía	Domicilio 2	NIE 2	1	2	G	823.2
.	.	.	.	.	.	.	.
Objeto n	Z S.A.	Domicilio n	NIE n	56	38	F	411.7

[2] Medición: valor medido para una variable de un objeto específico.
[3] Variable: característica medible de un objeto.
[4] Observación: todas las mediciones para un objeto específico, también se le conoce como entrada.

Al procesar las matrices de datos que aparecen en los gráficos 3.1, 3.2 o 3.3, es posible crear tablas como la siguiente. Agrupando los valores de las variables, se crean *variables agrupadas* y *variables divididas por intervalo de clase*.

En la tabla bidireccional que aparece abajo, hay dos *variables de expansión*: *rama industrial* y el *número de empleados*. La variable *rama industrial* se agrupó y la variable *número de empleados* se dividió en intervalos de clase.

Gráfico 3.4 Ejemplo de tabla creada con datos de los gráficos 3.1, 3.2 o 3.3

Empresas por rama industrial y número de empleados, 2000
Número de empresas

	Número de empleados					
Rama industrial	0	1–9	10–49	50–99	100–199	200–
A–F, bienes primarios	198 006	45 124	8 936	1 144	587	626
G–K, servicios privados	231 910	107 296	14 830	1 424	645	568
L–Q, servicios del sector público	65 783	24 621	3 713	434	200	548
No especificado	105 823	2 515	0	0	0	0

Las celdas de la tabla se definen combinando las variables de expansión. Las observaciones se distribuyen de forma que cada observación corresponda a una sola celda de la tabla. El valor en la celda de una tabla describe de manera sucinta las observaciones de la celda. Estas descripciones sucintas de grupos de observaciones se denominan *datos agregados* o *macrodatos*. Al usar la misma matriz de datos, es posible formar varios tabulados con diferente contenido.

Una matriz de datos se puede guardar en una tabla de base de datos, a la que a menudo se le llama tabla. Por desgracia, esto significa que sólo hay un término para dos conceptos diferentes. Usaremos el término *tabla* únicamente para referirnos al concepto estadístico.

3.2.2 Base de datos y tabla de base de datos

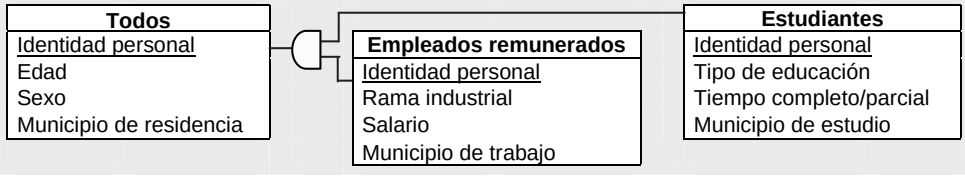
Las matrices de datos como las que aparecen en los gráficos 3.1 a 3.3 se guardan en bases de datos, que usan varios términos informáticos que describiremos en esta sección. En la sección 2.1 se describen modelos conceptuales, que a menudo difieren del establecimiento físico de una base de datos. En este libro, siempre hablamos en un nivel conceptual, es decir, cómo se *piensa* que se "verán" los registros y las matrices de datos. Por ejemplo, en el gráfico 3.1 hay una concordancia completa entre el registro, la matriz de datos y la tabla de base de datos, pero quizá no ocurra así con una solución informática real.

El objetivo de una solución de sistemas es que la base de datos sea flexible y pueda tener usos múltiples en el sistema de registros. La estructura de la matriz de datos también se puede guardar en una "vista". En el capítulo 11 veremos términos informáticos como *base de datos*, *tabla de base de datos* y *vista*.

En el gráfico 3.5 se presentan distintas formas de describir la misma base de datos. La parte A muestra la manera tradicional de describir una base de datos y la parte B un ejemplo del contenido de una base de datos. El gráfico ilustra la estructura de la solución de sistemas, es decir, la manera en que los datos se almacenan físicamente. La base de datos se normaliza para asegurar una consistencia óptima y un almacenamiento con uso eficaz del espacio.

Gráfico 3.5 Base de datos sobre personas con tres tablas de bases de datos

A. Modelo conceptual de la base de datos



B. Ejemplo de contenido en la base de datos

Todos

Id persona	Edad	Sexo	Municipio de residencia
NIP1	20	F	0586
NIP2	23	M	0586
NIP3	31	M	0586
NIP4	32	F	0586
NIP5	33	M	0586
NIP6	40	F	0586
NIP7	59	F	0586
NIP8	65	M	0586
NIP9	71	F	0586

Empleados remunerados

Id persona	Rama industrial	Salario	Municipio de trabajo
NIP2	G	52 000	0586
NIP3	G	287 000	0580
NIP4	A	193 000	0586
NIP6	D	291 000	0586
NIP7	D	314 000	0580

Estudiantes

Id persona	Tipo de educación	Tiempo completo/ parcial	Municipio de estudio
NIP1	Adultos	100	0586
NIP2	Univ	100	0580
NIP5	Univ	100	0580

Con la base de datos anterior, es posible llevar a cabo un procesamiento estadístico para distintas finalidades. En el gráfico 3.6 vemos un ejemplo de dos matrices de datos diferentes que podrían crearse a partir de esta base de datos. Estas matrices pueden generarse y almacenarse físicamente, pero también pueden generarse sólo de manera temporal durante el procesamiento. Una matriz de datos conceptual se centra en el problema específico que se desea resolver, es decir, el análisis estadístico de que se trate.

En el gráfico 3.6 tenemos un ejemplo de dos matrices de datos para empleo y traslado entre el lugar de residencia y el lugar de trabajo. En la terminología de las bases de datos, las matrices de datos del gráfico se denominarían *no normalizadas* porque varias de las celdas son valores perdidos.

Gráfico 3.6 Dos matrices de datos para diferentes propósitos estadísticos

A. Matriz de datos: Registro de empleo

Persona	Edad	Sexo	Empleado	Rama industrial	Salario
NIP1	20	F	No	nulo	0
NIP2	23	M	Sí	G	52 000
NIP3	31	M	Sí	G	287 000
NIP4	32	F	Sí	A	193 000
NIP5	33	M	No	nulo	0
NIP6	40	F	Sí	D	291 000
NIP7	59	F	Sí	D	314 000
NIP8	65	M	No	nulo	0
NIP9	71	F	No	nulo	0

B. Matriz de datos: Registro de traslado del lugar de residencia al lugar de trabajo

Persona	Municipio de residencia	Municipio de trabajo	Municipio de estudio	Traslado
NIP1	0586	nulo	0586	0
NIP2	0586	0586	0580	1
NIP3	0586	0580	nulo	1
NIP4	0586	0586	nulo	0
NIP5	0586	nulo	0580	1
NIP6	0586	0586	nulo	0
NIP7	0586	0580	nulo	1
NIP8	0586	nulo	nulo	0
NIP9	0586	nulo	nulo	0

A menudo se usa el término informático *tabla* en vez del término completo *tabla de base de datos* (que en realidad es una tabla relacional). Como en estadística el término *tabla* (o *tabulado*) se refiere a un método de presentación de datos agregados —es decir, *macrodatos*—, esa elección de vocabulario puede dar pie a malentendidos cuando se habla con estadísticos o con especialistas en informática. Las tablas de base de datos no deben confundirse con las tablas o tabulados estadísticos.

¿Qué queremos decir con el término *registro?* El ejemplo de los gráficos 3.5 y 3.6 se refiere a poblaciones completas, no muestras. Se incluye la identidad del objeto, lo que significa que estos dos gráficos muestran registros con el sentido que nosotros damos al término. También podríamos decir que la base de datos del gráfico 3.5 es un registro, con varios subregistros. Cuando se documenta el registro, debe darse la solución de sistemas del gráfico 3.5, más las matrices de datos estadísticas y relacionadas por contenido del gráfico 3.6. A continuación presentamos una síntesis de los términos *registro* y *matriz de datos*.

Gráfico 3.7 ¿A qué nos referimos con los términos *registro* y *matriz de datos*?

Registro	***Matriz de datos***
Son todos los objetos pertenecientes a un conjunto de objetos o población específicos definidos. A menudo consta de muchos subregistros y versiones de registro. Sus objetos son identificables. Contiene varios tipos de variables, incluidas las de tipo estadístico. Se almacena en una o varias bases de datos.	Es un conjunto de microdatos para uso en un procesamiento estadístico específico. Contiene principalmente variables estadísticas. Se puede anonimizar. Puede proceder de una encuesta por muestreo, un censo o una encuesta basada en registros. Se puede almacenar en una tabla de base de datos, pero también se puede crear en el momento del cálculo.

La terminología también debe ser comprensible para el público en general

Dentro de una oficina de estadística, en ocasiones el término *registro* se usa con diferentes significados. El término *tabla* también se relaciona algunas veces con una tabla de base de datos y otras con una tabla estadística o tabulado. Sin embargo, los términos usados externamente, en sitios web y publicaciones, deben ser comprensibles para el público en general.

Quizás el público en general interprete el término *registro* como toda la población —por ejemplo, un registro de miembros contiene a todos los miembros, no sólo una muestra—. También es probable que los usuarios del sitio web de una oficina de estadística relacionen el término *tabla* con una tabla estadística. Por consiguiente, en las comunicaciones externas los términos *registro* y *tabla* se deben usar siempre de una manera congruente y fácil de entender para los usuarios.

3.2.3 Términos para diferentes tipos de registros

La oficina de estadística recibe *registros administrativos* y los procesa para crear *registros estadísticos*. Según se describe en la sección 2.2, los cuatro registros estadísticos más importantes para el sistema de registros reciben el nombre de *registros base*, término que sólo se debe usar para esos cuatro registros. Los registros base se apoyan en fuentes administrativas, a partir de las cuales se generan los *conjuntos de objetos* y los *tipos de objetos* del sistema de registros.

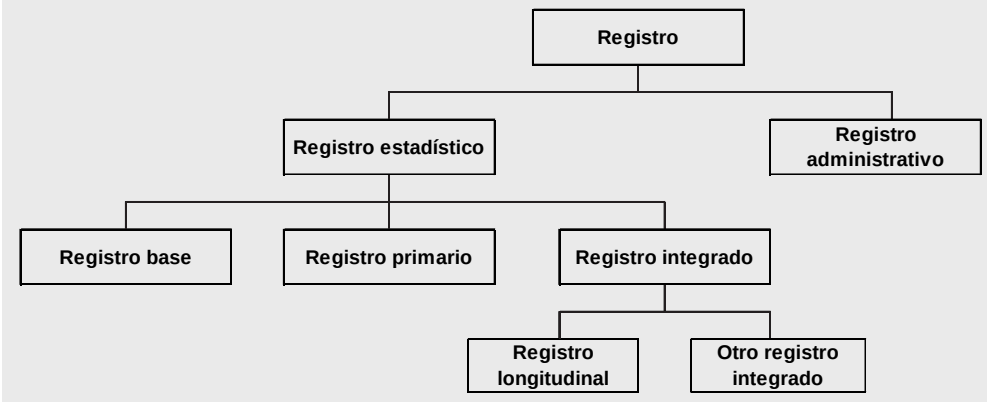
Los otros registros estadísticos (excepto los registros base) se dividen en dos categorías: *registros primarios* y *registros integrados*. Llamamos *registros primarios* a los registros estadísticos que se basan *directamente* en al menos una fuente administrativa. A partir de las fuentes administrativas, se genera la mayor parte de las *variables estadísticas* del sistema de registros.

Los *registros integrados* son registros estadísticos generados *exclusivamente* combinando información que ya existía en los registros estadísticos del sistema. Este término resalta una importante aplicación del sistema de registros: llevar a cabo nuevas encuestas usando sólo datos ya existentes.

Se puede combinar información de varios registros para crear un nuevo registro sin que éste sea un registro integrado. Por ejemplo, el Registro de Ingresos y Tributación se crea integrando seis fuentes administrativas y cinco registros estadísticos (gráfico 1.4). Sin embargo, es un registro primario porque se basa directamente en al menos una fuente administrativa.

Un *registro longitudinal* es un tipo especial de registro integrado donde se compila la información de varios registros anuales de modo que sea posible dar seguimiento a objetos idénticos a lo largo del tiempo.

Gráfico 3.8 Diferentes tipos de registros



3.2.4 Registros y tiempo

Los objetos individuales y, por tanto, los conjuntos de objetos cambian con el paso del tiempo. Los objetos nacen, cambian de ubicación, sufren modificaciones o dejan de existir. Estos diferentes tipos de ocurrencias, que reciben el nombre de *eventos demográficos*, son los que modifican los conjuntos de objetos. Cuando se define un registro en relación con el *tiempo*, deben distinguirse los siguientes tipos de registro:

1. El *registro de inventario actual* es la versión del registro actualizada con toda la información disponible sobre los objetos activos o vivientes en el momento. Se usa como población marco para encuestas por muestreo o censos.
2. El *registro referido a un momento específico*, como el final de un año, es la versión del registro actualizada para describir el conjunto de objetos en ese momento. Esta actualización se lleva a cabo *después* de ese momento, cuando está disponible la información sobre todos los eventos hasta entonces. Se usa para encuestas basadas en registros.
3. El *registro por año calendario* es la versión del registro que contiene todos los objetos que han existido en cualquier momento durante un año específico. Incluye los objetos

que se agregan o dejan de existir durante el año junto con información sobre la fecha del evento. Se usa como población de registro para encuestas basadas en registros.

Gráfico 3.9 Registro por año calendario para 2002

Identidad del objeto	Existía 1/1	Se agregó	Dejó de existir	Existía 31/12	Otras variables
Núm. de identificación 1	Sí	-	20020517	No	…
Núm. de identificación 2	Sí	-	-	Sí	…
Núm. de identificación 3	No	20020315	20020925	No	…
Núm. de identificación 4	No	20020606	-	Sí	…

4. El *registro de eventos* para un periodo específico es el registro que contiene información sobre todas las ocurrencias demográficas durante el periodo. Se crea un registro por cada tipo de evento. Se usa en las encuestas basadas en registros.

Gráfico 3.10 Registro de eventos para 2002 sobre cambio de domicilio

Identidad del objeto	Domicilio 1/1	Fecha de cambio de domicilio	Nuevo domicilio
Núm. de identificación 1	Domicilio 11	20020517	Domicilio 21
Núm. de identificación 2	Domicilio 12	20020606	Domicilio 22
Núm. de identificación 3	Domicilio 13	20020911	Domicilio 23

5. El *registro histórico* contiene información sobre todos los eventos demográficos que han ocurrido en cualquier momento para cada objeto. Un objeto que tiene, por ejemplo, tres eventos aparecerá en tres renglones de la matriz de datos. Se usa para encuestas longitudinales.

Gráfico 3.11 Registro histórico sobre cambio de domicilio

Identidad del objeto	Del domicilio	Fecha de cambio de domicilio	Al domicilio
Núm. de identificación 1	De nacimiento	19670517	Domicilio 1
Núm. de identificación 1	Domicilio 1	19810606	Domicilio 2
Núm. de identificación 1	Domicilio 2	20020911	Domicilio 3

6. Un *registro longitudinal* para un periodo (tres años en el gráfico 3.12) contiene información sobre eventos demográficos que ocurrieron durante el periodo en cuestión y los valores para las variables estadísticas de todos los subperiodos (en este caso, años) dentro de un periodo más largo.

Gráfico 3.12 Registro longitudinal para 2000–2002

Identidad del objeto	Existía 1/1/2000	Se agregó	Dejó de existir	Ingreso 2000	Ingreso 2001	Ingreso 2002
Núm. de identificación 1	Sí	-	20010517	183 450	97 600	-
Núm. de identificación 2	Sí	-	-	273 500	281 360	258 340
Núm. de identificación 3	No	20010315	20020925	-	193 570	204 520
Núm. de identificación 4	No	20020911	-	-	-	56 300

3.2.5 El término "objeto"

Hemos usado en muchas ocasiones el término *objeto*, con el que se relacionan varios conceptos y otros términos.

Población, objeto, tipo de objeto, instancia de objeto, clase, unidad, elemento
"La población *U* consta de *N* elementos/unidades/objetos". En la teoría de las encuestas, los análisis muchas veces se inician de esta manera. Casi siempre en los libros de texto de estadística se usan los términos *elemento* o *unidad*, pero el término *objeto* es común en las oficinas de estadística. El término correspondiente en la modelación conceptual de un sistema informático es *instancia de objeto*. Esta terminología se refiere a una población de individuos cuyo *tipo de objeto* es persona, donde cada individuo de la población es una instancia de objeto. A menudo se utiliza el término *clase de objeto*, sobre todo en la programación orientada a objetos, como sinónimo de tipo de objeto.

Para definir una población, deben definirse los tipos de objeto, es decir, qué se quiere decir con hogar, establecimiento, etc. También debe definirse el conjunto de objetos que se incluirá en la población y esa definición siempre contendrá información sobre lugar y tiempo, por ejemplo, hogares en cierto municipio en cierto momento.

El término* objeto *en estadística e informática
Dentro de las ciencias estadísticas reciben el nombre de *unidad* u *objeto* sólo aquellos objetos con los que se relaciona una encuesta estadística. En la sección 2.1 se presentan los diversos tipos de objeto que ocurren en las encuestas basadas en registros de una oficina de estadística:

- persona y hogar;
- organización, empresa y establecimiento;
- inmueble/propiedad y vivienda;
- vehículo;
- actividad, empleo remunerado y estudios.

En un ambiente informático, el término *objeto* se usa con frecuencia, a veces con definiciones que difieren de la acepción estadística. En una solución de base de datos, los renglones de algunas bases de datos reciben el nombre de objetos sin serlo en el sentido estadístico o conceptual. Esto puede ser motivo de confusión. Cuando se documenta una encuesta, sólo los objetos que forman parte de la población de registro deben llamarse objetos en la parte estadística de la documentación, mientras que la documentación referida al sistema informático puede contener otros objetos relacionados con el sistema.

3.3 TÉRMINOS PARA DIFERENTES TIPOS DE VARIABLES

Las columnas de una matriz de datos contienen mediciones para diferentes variables. Las variables pueden tener diferentes atributos y usarse de diferentes maneras en el procesamiento estadístico. Aquellas que forman parte de un sistema de registros pueden haber sido generadas de distintas formas y tener funciones especiales. Las variables son un tema que compete tanto a las ciencias estadísticas como a la informática.

3.3.1 Variables en las ciencias estadísticas

Una variable es un atributo medible de un objeto. Cuando se captan datos, recibimos valores que son mediciones de esos atributos. Lo que se compila en una matriz de datos son esas mediciones. En teoría, debemos diferenciar entre las características de un objeto y la medición recabada.

Una variable estadística está definida por el *tipo de objeto* que presenta la característica (por ejemplo, ingreso para personas e ingreso para hogares son dos variables distintas), por el *método de medición* y la *escala* aplicados, y por el *momento* o *periodo* a los que se refiere la medición.

Ejemplo: La característica "edad" de una persona en un momento específico puede medirse aplicando el criterio (joven, edad mediana o adulto mayor) o preguntando la fecha de nacimiento y calculando la edad.

Una variable puede ser *cuantitativa* o *cualitativa*. Las características que se pueden describir con un número (como la edad de una persona) se pueden medir con variables tanto cuantitativas (como edad en años) como cualitativas (como joven, de edad mediana o adulto mayor).

Las características correspondientes a una categoría (como sexo) sólo se pueden medir con variables *cualitativas*, aunque se usen códigos para denotar las diferentes categorías (por ejemplo, 1 para masculino y 2 para femenino). Una variable cualitativa está formada por la clasificación de objetos, es decir, cuando se dividen en varios grupos, motivo por el que estos grupos en ocasiones reciben el nombre de *clasificaciones*.

Algunas variables cualitativas contienen muchas categorías, que también se ordenan de manera *jerárquica*, como rama industrial para las empresas. La división es demasiado general en el nivel más alto de una clasificación jerárquica porque es de un solo dígito, se afina en el nivel de dos dígitos y así sucesivamente.

En el caso de las variables *cuantitativas*, los valores se pueden usar en cálculos y en una matriz de datos. Los valores anotados en una columna de variables se pueden agregar como sumas y medias. Con las variables cualitativas, es posible calcular el *número* con cierto código, por ejemplo, los valores 1 o 2 en la columna de la matriz de datos correspondiente a la variable "sexo".

Una variable *cualitativa* se puede transformar en variables cuantitativas con la codificación 0-1. Por ejemplo, el sexo se puede transformar en dos variables codificadas 0-1:

x_1 = 1 para femenino, 0 en otro caso x_2 = 1 para masculino, 0 en otro caso

x_1 se usa para describir el número de mujeres y x_2 para describir el número de hombres.

Se puede calcular tanto el valor medio como la desviación estándar para x_1 y x_2.

Variables, registros y tiempo

En la sección 3.2.4 veremos seis tipos de registro donde el tiempo se aborda de diferentes maneras. Hay dos tipos de variables donde el tiempo también recibe un tratamiento distinto. La combinación de tipo de variable y tipo de registro es importante y no debe pasarse por alto.

– Las *variables de flujo* muestran sumas para diferentes periodos, por ejemplo, el ingreso percibido durante un año por una persona o los nuevos pedidos recibidos durante un mes por una empresa.

– Las *variables de stock* presentan la situación en un momento específico, por ejemplo, la edad de una persona en determinado momento o el número de empleados de una empresa al final del año.

Se debe definir una variable de flujo para un registro por año calendario (o una versión del registro creada para un trimestre o un mes). Un registro del ingreso percibido de personas durante cierto año debe incluir a todas las personas pertenecientes a la población durante al menos una parte de ese año. La versión actual del Registro de Ingresos y Tributación de Suecia describe el ingreso durante el año *t* de las personas con residencia permanente en el país al 31 de diciembre del año *t*. Sin embargo, hay personas que residían permanentemente en Suecia antes del 31 de diciembre, pero que se fueron o murieron antes de esa fecha. En un registro por año calendario para el año *t*, se habrían incluido sus ingresos. En la versión actual del registro mencionado, el ingreso total es menor que en un registro por año calendario. De manera correspondiente, se debe definir una variable de *stock* para un registro referido a un momento específico.

3.3.2 Variables dentro de la informática

En el ámbito de la informática hay diversos términos para distinguir entre los diferentes tipos de variables. Por desgracia, estos términos normalmente no aparecen en la bibliografía sobre estadística; no obstante, como son importantes, en particular en un contexto estadístico basado en registros, nos referiremos a algunos de ellos.

El conjunto de valores que puede asumir, o puede pensarse que asuma, una variable para cualquier objeto se llama *conjunto de valores* de la variable. *Ejemplo:* La variable cuantitativa *edad de una persona* tiene un conjunto de valores "0–115 años".

Ejemplo: La variable cualitativa *condado de residencia* de una persona (en Suecia) tiene el conjunto de valores "condado de Estocolmo…, condado de Norrbotten", mientras que la variable cualitativa *código de condado* tiene el conjunto de valores "01, 03, …, 10, 12, 13, 14, 17, …, 25".

Una *variable de valor único* es aquella que sólo asume *un* valor para cada objeto. En general, las variables son de valor único. Ejemplo: La edad de una persona.

Una *variable de valores múltiples* puede asumir *varios* valores al mismo tiempo para ciertos objetos. El número de valores difiere de un objeto a otro. *Ejemplo:* La rama industrial de una empresa, pues una empresa puede tener actividades en varias ramas al mismo tiempo. El número de ramas industriales varía dependiendo de la empresa. Las variables de valores múltiples originan muchos problemas metodológicos relevantes que abordaremos en el capítulo 9.

3.3.3 Variables derivadas

Las variables derivadas desempeñan un papel central en las encuestas basadas en registros. Cuando recabamos datos en encuestas por muestreo o censos, diseñamos un cuestionario con las preguntas que definen las variables de la encuesta. En las encuestas basadas en registros, no tenemos esta oportunidad; en vez de diseñar cuestionarios, debemos crear variables derivadas usando todas las variables administrativas disponibles. También abordamos las variables derivadas en la sección 6.2.

Cuando procesamos datos estadísticos, buena parte del trabajo consiste en generar variables derivadas. Éstas se pueden formar a partir de variables definidas para los objetos incluidos en el registro en cuestión o usando variables definidas para otros objetos. En este último caso, será necesario emparejar diferentes registros. Se pueden distinguir cuatro tipos de variables derivadas:

1. Variables derivadas por agrupamiento de valores y división en intervalos de clase

NIP	Edad, años	Clase de edad	País de nacimiento	Categoría geográfica
1	76	70–79	Suecia	Suecia
2	49	40–49	Francia	Europa
3	32	30–39	Noruega	Países nórdicos
4	11	10–19	Chile	Fuera de Europa

Una variable *cuantitativa*, como edad, se puede dividir en *intervalos de clase*, por ejemplo, 0–9, 10–11, … 80–89, 90 y mayores.

Una variable *cualitativa*, como país de nacimiento, se puede *agrupar* en categorías más amplias, por ejemplo, Suecia, países nórdicos, Europa y fuera de Europa.

2. Variables derivadas mediante operaciones aritméticas usando variables de la matriz de datos

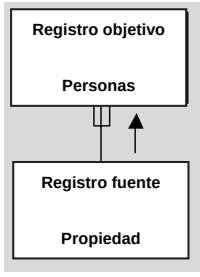
Con variables cuantitativas, es posible realizar operaciones aritméticas o lógicas, mientras que con variables cualitativas, sólo se pueden hacer operaciones lógicas.

Ejemplo: Ingreso disponible para personas = ingreso percibido + ingresos de capital + pagos de transferencia positivos – impuestos – pagos de transferencia negativos

Ejemplo: Número de unidades de consumo en un hogar. Definición sueca:
Un hogar con un adulto equivale a 0.2 • 1 + 0.96 • 1 = 1.16 unidades de consumo
Un hogar con dos adultos equivale a 0.96 • 2 = 1.92 unidades de consumo

Número de unidades de consumo en un hogar = 0.2 • (si sólo hay un adulto) + 0.96 • (número de adultos) + 0.76 • (número de niños entre 11–17) + 0.66 • (número de niños entre 4–10) + 0.56 • (número de niños entre 0–3) *Condición lógica subrayada = 1 si es verdad, 0 en caso contrario*

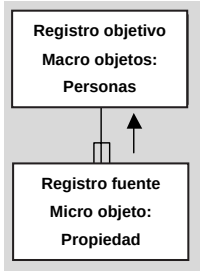
3. Variables derivadas por adjunción



Se genera una variable derivada en un registro usando variables de otro registro. Los objetos del primer registro se pueden vincular con los objetos del segundo registro en una relación *uno a uno* o *uno a muchos*. Esto significa que cada objeto del registro fuente se puede vincular con uno o más objetos del registro objetivo. A partir de esta relación, las variables del registro fuente se pueden adjuntar a los objetos del registro objetivo.

Ejemplo: En un registro sobre personas, se pueden adjuntar las coordenadas geográficas de la vivienda a cada persona. Los registros sobre personas pueden contener la identidad de la propiedad o la vivienda donde está registrada la persona. La identidad de la propiedad es el vínculo con el Registro Inmobiliario. Las coordenadas de la propiedad se transfieren del Registro Inmobiliario al registro correspondiente sobre personas. En este caso, las propiedades y las personas se vinculan en una relación *uno a muchos*, donde la propiedad se vincula con una o muchas personas.

4. Variables derivadas por agregación



Se crea una variable derivada en un registro usando variables de otro registro. Los objetos del registro fuente se pueden vincular con los objetos del registro objetivo mediante una relación *muchos a uno*. Uno o muchos objetos del registro fuente se pueden vincular con un objeto del registro objetivo. Es posible agregar valores para los *micro-objetos* del registro fuente que están vinculados con el *macro-objeto* respectivo en el registro objetivo, de manera que resulte relevante para la encuesta.

Ejemplo: El ingreso por hogar es una variable agregada que se forma sumando los valores de la variable *ingreso de la persona* para todas las personas de cierto hogar. En este caso, hogar es el macro-objeto y persona, el micro-objeto.

Ejemplo: ¿Cómo se puede combinar la información de un registro sobre personas con la de un registro sobre empresas? Para las empresas, se forma una variable derivada, *proporción de personas con educación superior*. Esta variable para el macro-objeto *empresa* se forma calculando la proporción de personas con educación superior entre todos los empleados (los micro-objetos) de la empresa.

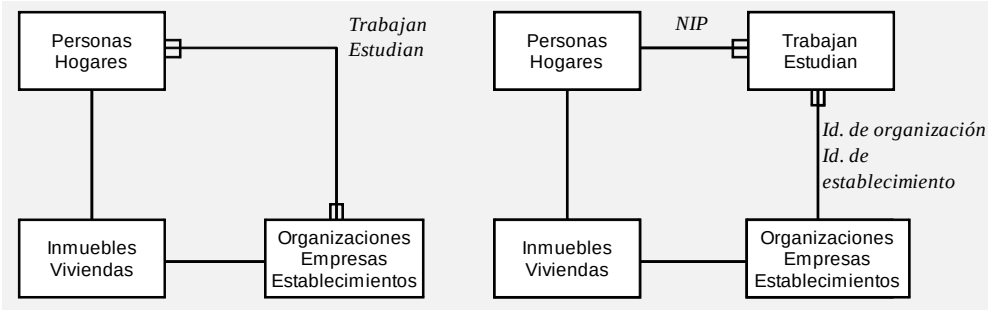
Un no emparejamiento generará valores perdidos en las variables derivadas tanto para las variables adjuntas como para las agregadas. En un sistema de registro eficaz con un bajo número de no emparejamientos, hay buenas posibilidades de formar variables adjuntas y agregadas de interés estadístico.

3.3.4 Adjunción, agregación y estructura del sistema de registros

Cuando generamos variables adjuntas o agregadas, estamos emparejando registros con diferentes objetos que guardan cierta clase de relación entre sí. Por medio de esta relación, las variables para una clase de objetos se transforman en variables para otra clase de objetos. Cuando la relación es *uno a uno*, esta transformación es simple; cuando la relación es *uno a muchos* o *muchos a uno*, la transformación se da por adjunción o agregación.

La relación *muchos a muchos* es complicada y debe evitarse. En el capítulo 2, analizamos la estructura del sistema de registros. El primer modelo conceptual del sistema se muestra en el gráfico 2.2, donde se ilustran las relaciones entre diferentes tipos de objetos. En el segundo modelo conceptual, gráfico 2.3, presentamos el Registro de Actividades como cuarto registro base. El contenido de estos dos gráficos aparece en el gráfico 3.13, donde indicamos con tridentes si las relaciones son *muchos a muchos*, *uno a muchos* o *muchos a uno*. Al introducir el Registro de Actividades con *Trabajan* y *Estudian* en el modelo, se evitan las relaciones *muchos a muchos* y los procesos de emparejamiento son más fáciles de entender.

Gráfico 3.13 Se deben evitar las relaciones muchos a muchos



Integración de registros con diferentes tipos de objetos

La adjunción y la agregación pueden explicarse mediante el ejemplo siguiente, donde empezamos con tres registros antes de llevar a cabo cualquier emparejamiento o creación de variables derivadas. Una persona puede tener varios empleos y un establecimiento, varios empleados.

Suma salarial es el término usado para tres variables diferentes:

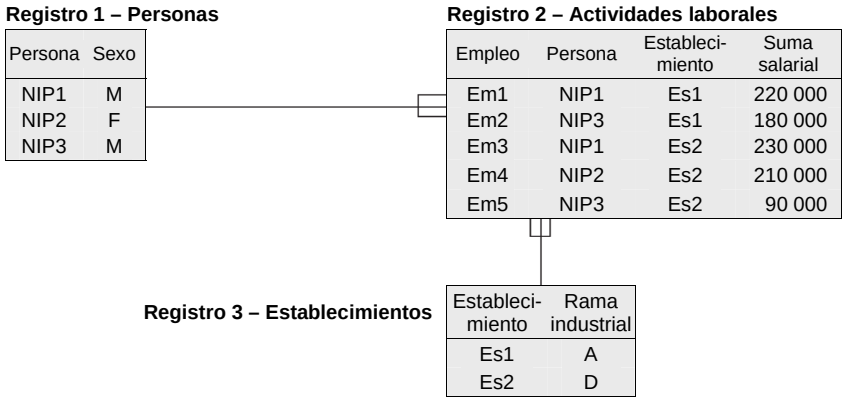
– Suma salarial para *empleo*, pago anual bruto por el trabajo que desempeña una persona en un empleo.

– Suma salarial para *persona*, pago anual bruto agregado por todos los empleos de una persona.

– Suma salarial para *establecimiento*, pago anual bruto agregado por todos los empleos en un establecimiento.

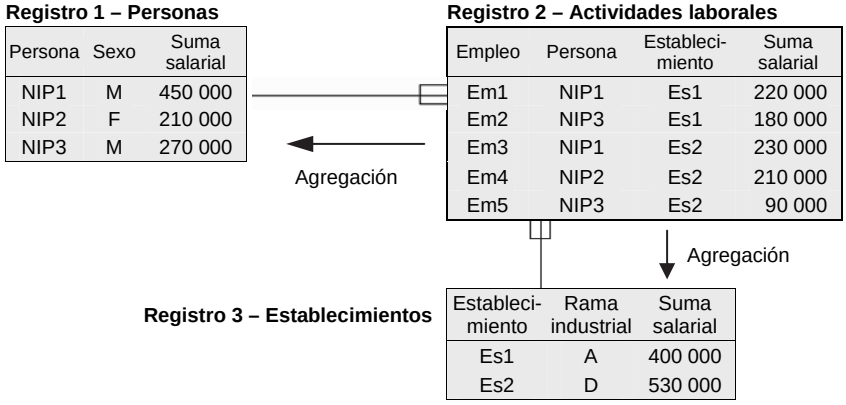
El ejemplo ilustra cómo se integran los datos de los tres registros con diferentes tipos de objetos. Se crean diversas variables derivadas mediante adjunción y agregación.

Gráfico 3. 14A Relaciones entre personas, actividades y establecimientos



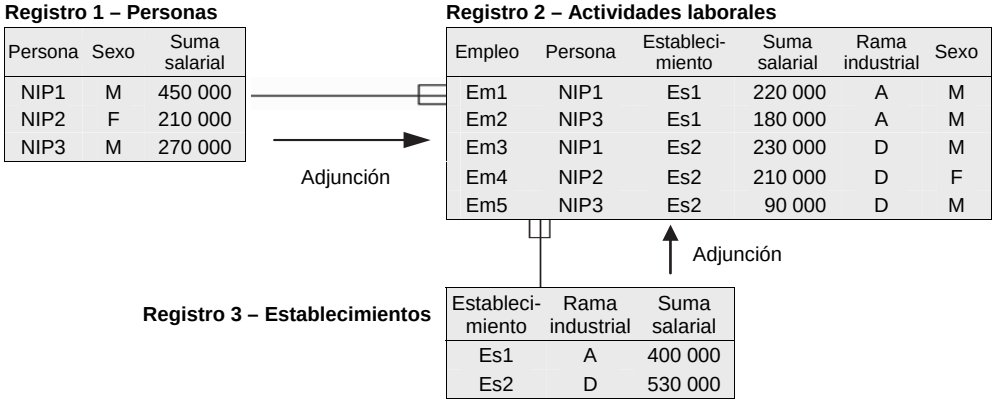
En la primera fase del proceso, ilustrado en el gráfico siguiente, las sumas salariales para las personas y los establecimientos se derivan por agregación de las sumas salariales del empleo. Los datos para empleo se agregan en un valor para cada persona o establecimiento. En el gráfico 3.14B hay tres variables distintas llamadas *suma salarial* definidas para tres diferentes tipos de objeto: personas, actividades laborales y establecimientos.

Gráfico 3.14B Sumas salariales para personas y establecimientos creados por agregación



En la segunda fase, gráfico 3.14C, podemos crear variables derivadas para las actividades laborales del Registro 2 adjuntando los valores de variables de los registros 1 y 3.

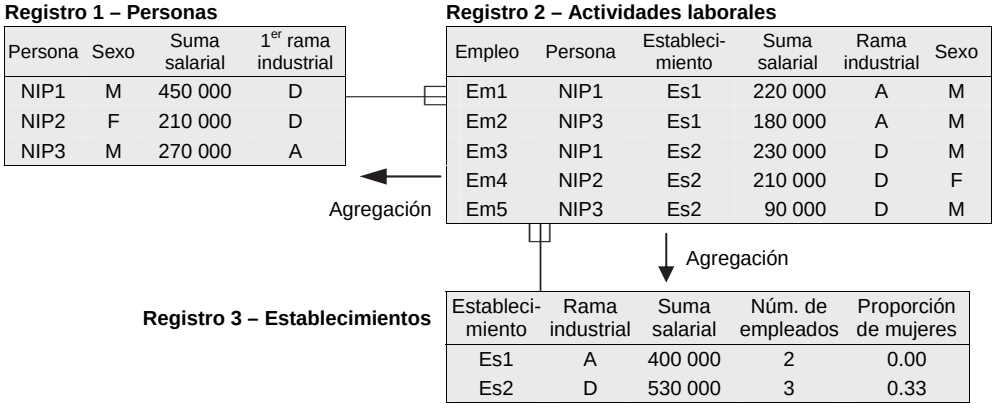
Gráfico 3.14C Rama industrial y sexo como variables derivadas para empleos generadas por adjunción



En la tercera fase, gráfico 3.14D, podemos crear más variables derivadas mediante la agregación de rama industrial y sexo del Registro 2. La variable *número de empleados* en el Registro 3 se creó contando los empleos del Registro 2 y la proporción de mujeres se obtuvo contando las mujeres de ese mismo registro y dividiendo el resultado entre el número de empleados.

Hasta ahora, todas las agregaciones han sido directas, pero cuando la rama industrial para actividades laborales se agrega a la rama industrial para personas, se vuelve más complicada la agregación de la variable cualitativa "rama industrial". Por cada persona del Registro 1, se calcula el tipo de valor de la rama industrial en el Registro 2 con las sumas salariales de ese registro como ponderadores o frecuencias.

Gráfico 3.14D Rama industrial, número de empleados y proporción de mujeres como variables derivadas, por agregación



En la tercera fase, se generaron inconsistencias en este sistema de tres registros. El número total de empleados es *tres* en el Registro 1 y *cinco* en el Registro 3. Las sumas salariales por rama industrial en el Registro 1 difieren de las sumas salariales por rama industrial en los otros registros.

La conclusión es que emparejar y derivar variables puede originar problemas metodológicos complejos. En el capítulo 9 analizaremos estos problemas.

3.3.5 Variables en el sistema de registros

En esta sección, describimos las variables del sistema de registros, en primer lugar en relación con el origen de las variables y en segundo, con su función en el sistema.

Variables con diferentes orígenes

Una variable puede ser de origen local o una variable puede ser importada. Una variable puede ser variable primaria o una variable derivada. Al combinar estos dos conceptos, obtenemos cuatro tipos de variables de origen diferente:

1. Variables tomadas directamente de un registro administrativo. Pueden haberse sometido a cierto procesamiento (el formato y los valores de las variables pueden haber cambiado luego de la validación). Estas variables reciben el nombre de *variables primarias locales*.
2. Variables derivadas que se crearon en el registro relevante. Reciben el nombre de *variables derivadas localmente*. Pertenecen a esta categoría todas las clases de variables derivadas como adjuntas y agregadas.
3. Variables tomadas de las variables primarias de otro registro estadístico con objetos idénticos. Se denominan *variables primarias importadas*.
4. Variables tomadas de variables derivadas de otro registro estadístico con objetos idénticos. Reciben el nombre de *variables derivadas importadas*.

Estos conceptos son útiles cuando el registro está documentado. Para un registro específico, sólo es necesario documentar las variables locales. La documentación de las variables importadas simplemente se debe transferir cuando se importan. Puede aplicarse el término *variables estandarizadas* a ciertas variables locales importantes para muchos registros. Estas últimas variables deben cumplir con estrictos requisitos adicionales en cuanto a documentación y calidad.

Gráfico 3.15 Cuadro simplificado del Registro de Ingresos y Tributación sueco para personas físicas

Variables primarias locales				*Variables derivadas locales*			*Variables importadas*	
NIP	**Ingreso1**	**.**	**Ingreso200**	**Ingreso derivado1**	**.**	**Ingreso disponible**	**Nivel de escolaridad máximo**	**Condado de residencia**
1	10 923	.	5 223	25 766	.	197 870	5	01
2	2 344	.	0	2 344	.	23 411	0	05
.	.	.	.	.	.	.	.	.
N	73 678	.	4 311	112 973	.	213 560	6	17

Las variables primarias locales de la matriz de datos anterior se basan en los datos administrativos que el Consejo Nacional Tributario proporciona al Registro de Ingresos y Tributación. Esas variables pueden servir para formar variables derivadas localmente. Se importaron otras dos variables: *nivel de escolaridad máximo*, variable derivada del Registro Educativo, y *condado de residencia*, variable primaria del Registro de Población que identifica el condado donde reside actualmente la persona. Originalmente, el nivel de escolaridad máximo que aparece en el gráfico 3.15 es una variable de valores múltiples, pues una persona puede tener, por ejemplo, dos doctorados. Aplicando la regla de transformación, esta variable se vuelve una variable de valor único con la definición “título obtenido más recientemente”.

Variables con diferentes funciones en el sistema

Distinguimos entre seis tipos de variables, cada una con una función distinta en el sistema de registros:

1. Las *variables identificadoras*, como el número de identificación, se usan, como su nombre lo indica, para identificar objetos. El término informático correspondiente es *clave primaria*. De ser posible, una variable identificadora debe ser completamente estable, es decir, tener el mismo valor durante toda la vida del objeto. Las variables identificadoras se usan cuando los registros que contienen el mismo tipo de objeto se emparejan para hallar coincidencias entre objetos idénticos. También se pueden usar variables como nombre, domicilio, etc., pero son más problemáticas para el emparejamiento. Por consiguiente, es preferible usar números de identidad cuando se procesan registros.
2. Las *variables de comunicación*, como nombre, domicilio y teléfono, se usan cuando la oficina de estadística necesita ponerse en contacto con un objeto en relación con un cuestionario o entrevista.
3. Las *variables de referencia (claves externas)* se usan para describir relaciones entre diferentes objetos. Cuando se emparejan registros que contienen datos sobre diferentes objetos, las variables de referencia generan coincidencias entre objetos relacionados.
4. Las *referencias temporales* son variables que indican el momento en que ocurre un evento, el cual afecta a los objetos o las actualizaciones del registro. Estas variables se usan cuando se generan las diferentes versiones del registro, como la población en un momento específico, y también para describir el flujo de los eventos demográficos durante un periodo.
5. Las *variables técnicas*, o variables para la *administración de registros* interna, a menudo indican la fuente o incluyen comentarios sobre temas o mediciones individuales. Por ejemplo, la fuente para el código de clasificación industrial de una empresa puede ser la Oficina de Patentes y Registro o el Consejo Nacional Tributario. También son de utilidad para mostrar los valores imputados, los códigos de corrección o los códigos de error. Las variables con ponderadores se usan para fines de estimación.
6. Las *variables estadísticas* reales se usan cuando se analiza y describe la matriz de datos. Las *variables de expansión* se utilizan para definir las celdas en las tablas estadísticas. Por cada celda de una tabla, se calculan las medidas descriptivas para otras variables estadísticas, las *variables de respuesta*.
   En ocasiones una variable estadística puede ser una variable de expansión en un contexto y una variable de respuesta en otro contexto. Por ejemplo, cuando se calcula el salario promedio para los distintos sexos, el sexo funciona como variable de expansión y el salario como variable de respuesta. Si se calcula la proporción de mujeres para diferentes ocupaciones, la ocupación será la variable de expansión y el sexo, la variable de respuesta. Los registros base deben contener variables de expansión relevantes para muchos usuarios, lo cual favorecerá la consistencia.

Las variables se deben documentar de diferentes maneras, dependiendo de cuál de estas seis funciones cumplan en un registro. Es importante que las variables estadísticas reales estén bien documentadas.

Variables usadas para emparejamiento

Debemos distinguir entre dos tipos de emparejamiento:

– El que tiene como propósito encontrar *objetos idénticos* en diferentes registros o tablas de base de datos. Durante el emparejamiento, una o más variables identificadoras presentes en los registros relevantes se usan como variables de vinculación.

Ejemplo: Se emparejan dos registros sobre personas; la variable de vinculación *número de identificación personal* estará presente en ambos.

– El que tiene como propósito encontrar objetos que guardan *cierto tipo de relación* entre sí. Estos objetos se pueden hallar dentro de la misma tabla de base de datos o en diferentes registros concernientes a tipos de objeto diferentes. Durante el emparejamiento, una variable de referencia del primer registro y una variable correspondiente del segundo registro se usan como variables de vinculación.
Ejemplo: Un registro sobre personas contiene la variable identificadora *número de identificación personal*, pero también una variable de referencia que proporciona el número de identificación personal del cónyuge de esa persona. Se emparejan dos copias de este registro usando el número de identificación personal como clave primaria en la primera copia y la variable de referencia como clave externa en la otra copia.
Ejemplo: Un registro sobre personas con número de identificación personal como clave primaria se puede emparejar con un registro de actividades en el que las actividades remuneradas se identifican con tres variables: número de identificación personal, número de establecimiento y número de organización. El número de identificación personal en el registro de actividades es la clave externa para el emparejamiento.

Cuando se emparejan dos registros, A y B, se crean tres nuevos conjuntos de objetos: emparejados AB, no emparejados de A y no emparejados de B. Estos tres conjuntos de objetos se deben guardar y documentar. Cuando se emparejan tres registros, A, B y C, se crean siete nuevos conjuntos de objetos que se deben documentar: emparejados ABC, emparejados en dos registros (AB, AC o BC) y no emparejados de A, B y C.

Gráfico 3.16 Combinación de conjuntos de objetos: interpretación del no emparejamiento

Cuando se combinan diferentes registros mediante emparejamiento, por lo general surgen diferentes tipos de no emparejamiento y debe decidirse cómo manejar las diferentes categorías de objetos. Asimismo, se debe definir qué categorías se incluirán en el nuevo registro y la causa del no emparejamiento.

Los emparejamientos se realizan por medio de *variables de vinculación*, que es un término que engloba tanto variables como variables de referencia. Un *vínculo* entre dos registros consta de una o más variables de vinculación. El término estadístico *vínculo* corresponde al término informático *clave*. Una variable identificadora puede ser una *clave primaria* y una variable de referencia, una *clave externa*.

Tipos de registro y de variable

Resumiendo lo planteado en los capítulos 2 y 3, podemos comparar distintos tipos de registro en el sistema de registros. Hay diferencias entre los diferentes registros en cuanto a los tipos de variable que son importantes. Además, los distintos tipos de registro pueden diferir en lo que respecta al tipo de trabajo al que están destinados y su función dentro del sistema de registros.

Gráfico 3.17 Función primaria de un registro en el sistema

Tipo de registro	Tipos de variables	Función y responsabilidad
Registro base	*Variables primarias locales:* - Variables identificadoras - Variables de comunicación - Variables de referencia - Referencias temporales	- Recibir datos administrativos - Generar conjuntos de objetos - Definir objetos - Generar algunas variables de expansión básicas - Producir estadísticas demográficas
Registro primario	*Variables primarias locales:* - Variables identificadoras - Variables estadísticas	- Recibir datos administrativos - Generar las variables estadísticas reales
Registro integrado	*Variables importadas:* - Variables identificadoras - Variables estadísticas - Variables derivadas localmente, variables adjuntas y agregadas	- Generar nueva información sin captación de datos - Compilar información de diferentes campos - Compilar información de diferentes periodos

CAPÍTULO 4

Encuestas por muestreo y registros

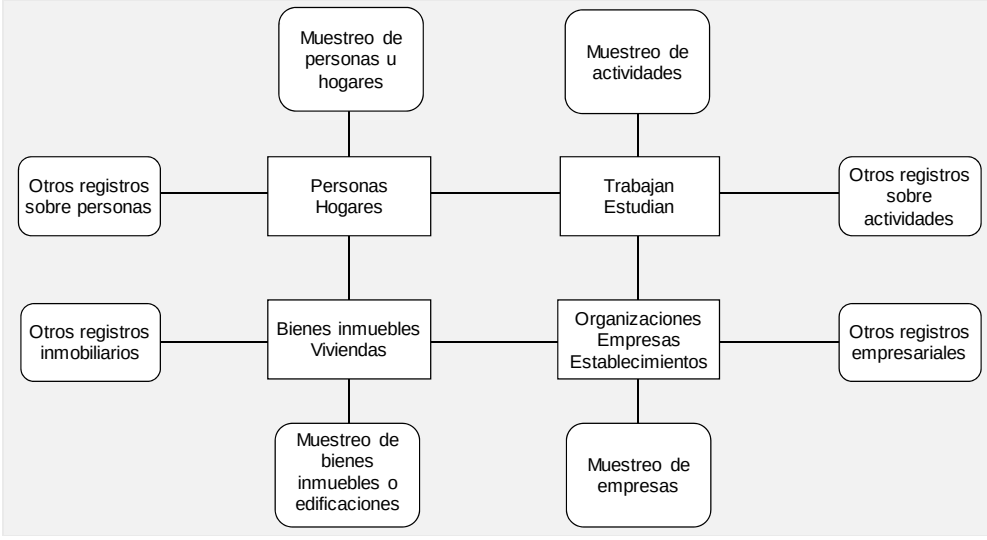
Una oficina de estadística puede llevar a cabo encuestas por muestreo y encuestas basadas en registros. Veremos que estos dos tipos de encuesta pueden beneficiarse entre sí y combinarse de diferentes maneras. Asimismo, compararemos las metodologías aplicadas a las encuestas por muestreo, los censos y las encuestas basadas en registros.

4.1 ¿EN QUÉ BENEFICIA UN SISTEMA DE REGISTROS A LAS ENCUESTAS POR MUESTREO?

Contar con un sistema de registros bien establecido repercute de manera importante en las posibilidades de hacer encuestas por muestreo. El sistema de registros puede usarse de las siguientes maneras:

– Al seleccionar la muestra, el registro base apropiado funciona como marco muestral y las variables del registro se usan para estratificar la población.
– Las mediciones se pueden facilitar eliminando la necesidad de hacer preguntas sobre datos ya disponibles en los registros.
– En la fase de estimación, las variables del registro se pueden usar como información auxiliar para aumentar la precisión y compensar la no respuesta.

Gráfico 4.1 El sistema de registros y el sistema de encuestas por muestreo



Muestreo

Los registros base funcionan como marcos de muestreo y las variables de los registros se usan para producir muestras estratificadas. Dentro de los estratos, como regla general es aplicado el muestreo aleatorio simple. Los registros base contienen identidades y domicilios de objetos pertenecientes a la población marco. Mediante estas identidades, los objetos de cada muestra se pueden vincular con otros registros del sistema y las variables de los registros se pueden importar a la matriz de datos de una muestra. En el gráfico 4.1 se ilustra la estructura de todo el sistema estadístico, donde las encuestas por muestreo se agregan al sistema de registros estadísticos.

Ejemplo: En la Encuesta sobre Fuerza Laboral de la Oficina de Estadística de Suecia, se usa el Registro de Población para seleccionar una muestra entre la población de 16 a 64 años que se estratifica por región, sexo y ciudadanía de acuerdo con ese registro, así como por trabajadores asalariados y no asalariados de acuerdo con el Registro de Empleo. El nivel de escolaridad de cada entrevistado se importa del Registro de Educación y el código industrial del establecimiento donde trabaja el informante se importa del Registro Empresarial. Los códigos regionales se importan de la base de datos geográfica.

Captación de datos y mediciones

Como no es necesario preguntar datos que ya existen en los registros de la Oficina de Estadística de Suecia, se reduce la carga para los informantes, en parte porque disminuye el tiempo de la entrevista y en parte porque los entrevistadores no necesitan hacer preguntas delicadas y difíciles, por ejemplo, sobre ingresos, edad y nivel de escolaridad.

Ejemplo: La Encuesta sobre Distribución del Ingreso usa sobre todo variables de ingreso del Registro de Ingresos. Las personas seleccionadas sólo deben responder las preguntas sobre la composición del hogar, si trabajan de tiempo completo o medio tiempo y sobre ciertos tipos de ingresos no asentados en el registro.

Otra manera de combinar una encuesta por muestreo con datos de registros consiste en abordar la situación actual mediante un cuestionario o entrevista y usar los datos de registro para retratar la situación histórica.

Ejemplo: En una encuesta sobre salud, se hacen preguntas sobre el estado de salud y las condiciones de vida actuales, mientras que los datos de registros pueden servir para ilustrar la formación académica, los antecedentes laborales, etc.

Estimación

Los registros también desempeñan un papel importante en la fase de estimación. El sistema de registros contiene muchas variables que pueden usarse como información auxiliar para producir mejores estimaciones. Estas variables auxiliares minimizan el error estándar de las estimaciones con un tamaño de muestra fijo o reducen el tamaño de la muestra y el costo de las estimaciones con un estándar de exactitud fijo.

Ejemplo: En la fase de estimación de la Encuesta sobre Fuerza Laboral, las variables de registro *sexo* y *edad* se toman del Registro de Población. El empleo remunerado por rama industrial se importa del Registro de Empleo y la categoría de quienes buscan empleo, del registro sobre este tema que mantiene el Consejo Sueco del Mercado de Trabajo. De acuerdo con Hörngren (1992), es razonable considerar que estas variables auxiliares reducen los márgenes de error aproximadamente 20% para un tamaño de muestra fijo, lo que corresponde a un tamaño de muestra reducido de alrededor de 36% para el mismo estándar de exactitud fijo.

La no respuesta en las encuestas por muestreo es un problema importante y su magnitud va en aumento. Los datos de los registros se pueden usar para *reducir la magnitud* de la no

respuesta porque permiten omitir preguntas sobre ciertas variables sensibles como ingresos, nivel de educación y edad. Asimismo, los cuestionarios se abrevian si se recurre a variables de los registros en vez de preguntas. Esto también puede ayudar a disminuir la magnitud de la no respuesta.

Las variables de los registros también pueden servir para minimizar los efectos del *error por no respuesta*. Por consiguiente, resulta conveniente tener acceso a muchas variables porque esto permite elegir el método de ajuste por no respuesta más adecuado para la encuesta de que se trate. Särndal y Lundström (2005) ofrecen una descripción completa del uso de las variables de registros para calibrar los ponderadores de muestreo y así reducir los errores de muestreo, los errores por no respuesta y los errores en el marco.

4.2 COMBINACIÓN DE ENCUESTAS BASADAS EN REGISTROS Y ENCUESTAS POR MUESTREO

Las encuestas basadas en registros y las encuestas por muestreo u otros tipos de captación de datos se pueden combinar de diversas maneras. A continuación explicamos brevemente algunos casos en los que es importante combinar estos dos métodos.

1. Definición de una población objetivo precisa

Una encuesta basada en registros y una encuesta por muestreo pueden complementarse en cuanto al contenido. Las estadísticas basadas en registros proporcionan datos básicos sobre las diferencias entre categorías y cambios en el transcurso del tiempo. Si también se lleva a cabo una encuesta por muestreo, obtendremos mayores detalles sobre el porqué de estas trayectorias. Cuando se usan registros, la muestra se puede limitar a subgrupos específicos de interés.

Ejemplo: Si se usa el Registro sobre el Impuesto al Valor Agregado, se puede seleccionar una muestra de las empresas (unidades jurídicas) que tuvieron una menor facturación en el año anterior. Después se puede entrevistar a estas empresas seleccionadas sobre los motivos de esa disminución y sus predicciones para el futuro.

Ejemplo: Si se usa el Registro de Actividades, se puede seleccionar una muestra de personas que han cambiado de empleador en los últimos dos años y entrevistarlos sobre los motivos de esos cambios.

2. Encuestas por muestreo y calidad de los registros

La calidad de los registros base es de vital importancia para todo el sistema. Al analizar las encuestas por muestreo realizadas para otros propósitos, es posible identificar si hay una sobrecobertura. Una encuesta empresarial permite estimar la proporción de empresas inactivas en el Registro Empresarial. En el caso de encuestas de personas y hogares, la no respuesta en entrevistas telefónicas o las devoluciones de cuestionarios enviados por correo se pueden analizar para obtener indicios de sobrecobertura en el Registro de Población.

La calidad de las principales variables de clasificación, por ejemplo, del Registro Empresarial, se debe monitorear mediante encuestas por muestreo especiales para verificar la rama industrial y otras variables. Todas las encuestas por muestreo comunes con variables similares a las variables de los registros estadísticos se deben usar para sondear problemas de medición. Las causas de estos errores pueden encontrarse tanto en la encuesta por muestreo como en el registro. Volveremos a este tema en la sección 5.4.7.

3. Mejoramiento de la calidad de los registros mediante encuestas para mantenimiento de registros

Ciertas categorías de objetos en un registro base pueden ser imperfectas, pueden faltar variables importantes en los registros o se puede sospechar que sean de baja calidad. Los cuestionarios enviados por correo o las entrevistas telefónicas sobre esas categorías mejorarán la calidad del registro y deberán usarse regularmente como métodos para mantener un registro base de alta calidad.

4. Complementación de un registro administrativo con una captación de datos especial

Una fuente administrativa con variables estadísticamente importantes puede estar incompleta a causa de una subcobertura. Por ejemplo, en algunos sistemas administrativos pueden faltar empresas pequeñas. Si un registro administrativo de ese tipo se compara con un registro base apropiado, puede ocurrir que este último incluya una categoría de objetos que no se encuentra en el registro administrativo. En este caso, la población debe dividirse en dos estratos: la que está presente en la fuente administrativa y la que aparece en el registro base, pero no en la fuente administrativa. El primer estrato se estudia mediante una encuesta basada en registros y el segundo mediante una encuesta por muestreo o un censo. Selander *et al*. (1998) analizan este método.

5. Uso de encuestas por muestreo para generar variables derivadas en los registros

Las variables estadísticas en las encuestas por muestreo se pueden comparar con las variables administrativas de los registros. Es posible construir modelos de medición útiles para determinar cómo se deben usar las variables administrativas cuando se generan variables derivadas para propósitos estadísticos, como veremos en la sección 6.2.3.

6. Estimación de áreas pequeñas

"La estimación de áreas pequeñas de variables estudiadas en encuestas sociales es una necesidad cada vez más imperiosa para los gobiernos, sobre todo en lo que respecta al establecimiento de una mejor asignación de recursos dirigidos para problemas sanitarios, vivienda precaria, desempleo y bajos salarios". Ésta es la introducción de Heady *et al*. (2003) para su informe sobre estimación de áreas pequeñas, lo que pone de relieve la gran demanda de estadísticas que describan áreas geográficas pequeñas.

Heady *et al*. formulan métodos para la estimación de áreas pequeñas en Inglaterra y Gales. En los países escandinavos esta demanda se satisface mediante encuestas basadas en registros, pero en los países donde la oficina de estadística no tienen acceso a registros administrativos con un contenido amplio, que pueda servir para propósitos estadísticos, la experiencia en la estimación de áreas pequeñas revisten gran interés.

En el informe citado también se describen las acciones de largo plazo necesarias para fomentar la capacidad en este tipo de modelación estadística. Los métodos diseñados por sus autores se basan en un uso combinado de datos obtenidos en encuestas por muestreo y datos administrativos o datos de censos anteriores. Los microdatos de la muestra se enriquecen con datos agregados de los registros administrativos o de los censos. Entonces la observación para cada persona de la muestra contendrá datos de la entrevista o el cuestionario usados en la encuesta por muestreo y covariables en forma de registro agregado o datos censales para el área geográfica pequeña de esa persona. En los modelos de regresión de niveles múltiples, las covariables explican cada una de las variables de interés de la encuesta por muestreo. Estos modelos estimados, junto con los datos administrativos o censales agregados para todas las áreas geográficas pequeñas, se utilizan con el propósito de producir estimaciones para todas las áreas.

7. Censos virtuales

Un censo tradicional de población y vivienda se apoya por completo en la captación de datos. Por su parte, un censo basado en registros se apoya por completo en registros administrativos, pero requiere pleno acceso a fuentes administrativas con toda la información necesaria.

El término *censo virtual* se refiere a un censo basado en una combinación de registros administrativos y encuestas por muestreo. La Oficina de Estadística de los Países Bajos (Statistics Netherlands, 2004) describe el censo virtual neerlandés de 2001. Para un censo virtual se requieren algunas fuentes administrativas, como un registro de población, que se pueden complementar con encuestas por muestreo en caso de que falte información.

El censo virtual neerlandés de 2001 se basó completamente en fuentes que ya existían. Con este método se logró un ahorro sustancial: el costo de un censo tradicional habría ascendido a 300 millones de euros, mientras que el costo del censo virtual fue de apenas 3 millones de euros. Hubo otras razones para realizar un censo virtual, por ejemplo, la disposición para participar en un censo tradicional ha disminuido, mientras que un censo virtual goza de una mayor aceptación social. La participación en un censo tradicional sería limitada y selectiva, lo que dificultaría la corrección de errores por no respuesta.

Un aspecto metodológico destacado es la manera en que los datos de diferentes registros administrativos y encuestas por muestreo se pueden combinar para producir estimaciones consistentes que permitan elaborar los tabulados censales que se publicarán. Houbiers *et al*. (2003) describen el método de ponderación repetida, que se usa para ajustar las estimaciones con base en las encuestas por muestreo, de modo que se asegure la consistencia de todas las estimaciones publicadas.

Los métodos de estimación para áreas pequeñas también pueden ser útiles en un censo virtual. Esto se plantea como una posibilidad interesante para futuros censos virtuales en los Países Bajos. Cuando se disponga de más fuentes administrativas, se podrá remplazar un censo virtual por un censo basado en registros.

4.3 COMPARACIÓN DE ENCUESTAS POR MUESTREO Y ENCUESTAS BASADAS EN REGISTROS

La planeación de una encuesta basada en registros es completamente distinta de la de una encuesta por muestreo. En una encuesta por muestreo, el primer paso consiste en decidir cuál será la población y qué parámetros deberán estimarse para qué áreas de interés. Esto determina el carácter de la encuesta en lo que respecta al diseño y la estimación de la muestra. Primero se definen la población y los parámetros, seguidos de la captación de los datos. En general, se considera *una* encuesta que se realiza en un momento con un número limitado de parámetros y de áreas de interés. Cuando se trabaja con encuestas por muestreo a menudo se usan los pares de conceptos *población–muestra* y *parámetro–estimador*.

En una encuesta basada en registros, el punto de partida es muy distinto, pues los datos ya fueron captados y están disponibles en diferentes registros, sin estar adaptados para ninguna encuesta específica. Con base en los registros disponibles, se seleccionan los objetos y variables relevantes para la encuesta de que se trate. En ocasiones es necesario formar nuevas variables a partir de los registros existentes—algunas veces incluso nuevos objetos (unidades estadísticas) —. Pero partamos del supuesto de que los datos ya existen. Luego se define la población y, posteriormente, el plan de tabulados sin pensar en elementos como estimación de parámetros y áreas de interés. El error de muestreo no limita las posibilidades

de elegir áreas para el análisis y la presentación de resultados. En las estadísticas basadas en registros, la distinción entre *población* y *muestra* carece de relevancia y no es común el uso de los términos *parámetro* y *estimador*.

Resulta conveniente que el diseño de un registro estadístico y un sistema de registros sea flexible a fin de darles una aplicación lo más amplia posible. Por consiguiente, una parte importante de la metodología estadística para el aprovechamiento de registros consiste en estructurar y mejorar el conjunto, es decir, encontrar el mejor diseño posible para todo el sistema de registros. Esto incluye un trabajo de largo plazo en el monitoreo y control del acceso a datos administrativos para propósitos estadísticos.

Por lo tanto, aunque puede considerarse que gran parte de la metodología estadística se aplica lo mismo a las encuestas por muestreo que a las encuestas basadas en registros —por ejemplo, en caso de problemas ocasionados por errores (salvo el error de muestreo) y en el trabajo de análisis y presentación—, el planteamiento es distinto, habida cuenta de que los errores de muestreo y los problemas de diseño son cruciales para las encuestas por muestreo. En las estadísticas basadas en registros, es fundamental *pensar en función de un sistema*. Para mejorar la calidad, no basta con abordar un registro cada vez, sino que se debe considerar todo el sistema. Se debe prestar especial atención a la calidad de los registros base y las variables identificadoras que actúan como vínculos entre los diversos registros.

Se deben cumplir varios prerrequisitos antes de pasar al *proceso de validación* tanto de las encuestas por muestreo como de las encuestas basadas en registros. Si se detectan valores no razonables al verificar una encuesta por muestreo, es posible ponerse en contacto de nuevo con el informante correspondiente. También es posible repetir una pregunta durante una entrevista si una respuesta parece extraña. Estas soluciones no se aplican a las encuestas basadas en registros. Sin embargo, se pueden hacer preguntas a la autoridad que suministra el registro sobre qué debe considerarse un valor razonable y cómo pueden ocurrir errores. Es importante que la oficina de estadística sepa cuáles son las variables verificadas y corregidas por la autoridad.

Las encuestas basadas en registros también presentan problemas de validación en otras situaciones. Cuando se reciben registros grandes, quizá sea necesario verificar y corregir errores técnicos. Por ejemplo, los datos escaneados pueden tener valores en posiciones incorrectas. En ciertos casos, los datos no se recopilan de manera centralizada, sino que varias autoridades regionales se ocupan de ello y después, en la oficina de estadística, se investiga si hay diferencias estructurales entre las distintas autoridades respecto a retrasos, patrones de respuesta, etc.

Asimismo, el problema de la *no respuesta* es de índole distinta para las encuestas por muestreo y las encuestas basadas en registros. También hay valores perdidos en las encuestas basadas en registros, pero las personas con las que no se pudo establecer contacto o que se negaron a responder no causan el mismo problema que en las encuestas por muestreo o censos, y no es necesario enviar recordatorios o decidir cuándo se debe dar por concluida la captación de datos. En las encuestas por muestreo y los censos, las estimaciones se revisan para contrarrestar los efectos de la no respuesta. Esto no siempre ocurre con las estadísticas basadas en registros.

Los *errores de medición* existen tanto en las encuestas por muestreo como en las encuestas basadas en registros. En el primer caso, nos centramos en minimizar el error de medición poniendo a prueba y mejorando los cuestionarios. No podemos trabajar de la misma manera con las estadísticas basadas en registros, pues los cuestionarios están diseñados por autoridades administrativas. Sin embargo, una oficina nacional de estadística debe influir de

alguna manera en las distintas autoridades con el propósito de contribuir al mejoramiento de los formatos de captación y de las instrucciones de respuesta.

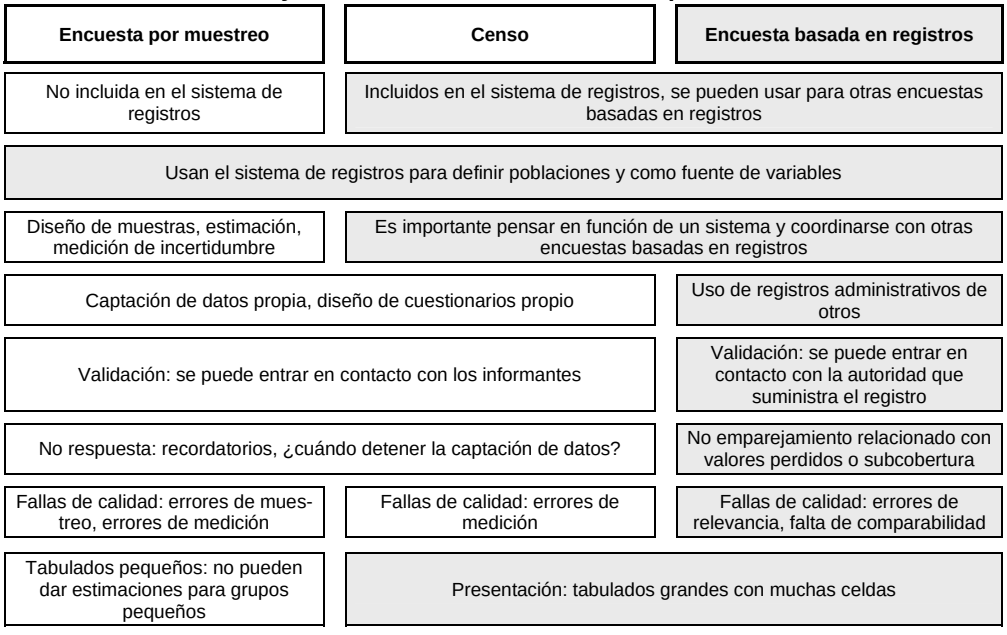
Los prerrequisitos para la *presentación* de los resultados de la encuesta difieren un poco. No podemos presentar resultados de encuestas por muestreo para grupos demasiado pequeños, considerando que los límites están definidos por el error de muestreo. Por consiguiente, los tabulados no pueden desglosarse en muchas variables a la vez, ni tener demasiadas celdas. En cambio, en las estadísticas basadas en registros, es posible producir grandes tabulados detallados que se desglosan de muchas maneras. A menudo estos tabulados son difíciles de interpretar y su método de presentación exige que se cumplan varios requisitos. Además, los *riesgos de exposición* para las distintas personas y empresas son otro factor que limita el nivel de detalle de los tabulados.

Como hemos visto, no es raro que las encuestas por muestreo y las encuestas basadas en registros planteen diferentes problemas metodológicos. A la fecha, la metodología para las áreas de validación, no respuesta y errores de medición se refiere principalmente a las encuestas por muestreo.

Para evitar confusiones con los diferentes conceptos, es importante distinguir con claridad entre los tres tipos de encuesta que puede realizar una oficina de estadística. Aquí comparamos las similitudes y las diferencias entre encuestas por muestreo, censos y encuestas basadas en registros. Estas últimas, que analizaremos con detalle en los siguientes capítulos, aparecen sombreadas en el gráfico 4.2.

Tanto los datos captados para censos como los registros basados en datos administrativos se pueden incluir en el sistema de registros si los datos contienen variables de identificación, de modo que los datos se puedan vincular con otros registros. Sin embargo, las encuestas por muestreo no forman parte del sistema de registros. Pueden utilizar los registros del sistema, pero no al revés: los datos de las encuestas por muestreo no pueden usarse en los registros del sistema, de modo que no se incluyen en el sistema de registros.

Gráfico 4.2 Similitudes y diferencias entre los distintos tipos de encuestas



CAPÍTULO 5

Cómo crear un registro: la población

La metodología es tan necesaria para las encuestas basadas en registros como lo es para las encuestas por muestreo. Cuando empezamos a trabajar en las estadísticas basadas en registros, observamos que muchas unidades de la Oficina de Estadística de Suecia habían llevado a cabo durante mucho tiempo un procesamiento de registros avanzado, lo que les permitió acumular un conocimiento considerable sobre cómo debían producirse las estadísticas basadas en registros. Obviamente, se aplicaban métodos, pero no existía una metodología general como tal. En este capítulo y el siguiente, formulamos algunos conceptos y principios generales que pueden servir como primer paso hacia el establecimiento de una metodología estadística para el aprovechamiento de registros.

La metodología tradicional sobre encuestas responde a la pregunta: *¿cómo se deben llevar a cabo las encuestas y censos?* De igual modo, la metodología estadística para registros debe responder a la pregunta: *¿cómo se debe llevar a cabo una encuesta basada en registros?*

Sin embargo, como ningún registro estadístico está diseñado exclusivamente para *una* encuesta específica, sino que debe servir para diferentes encuestas ya sea solo o en combinación con otros registros, los principios de la metodología deben ser más flexibles. Es decir, la metodología estadística para registros debe responder a la pregunta: *¿cómo se debe crear un registro estadístico de modo que pueda servir no sólo para una encuesta específica, sino también contribuir a la realización de otras encuestas dentro del sistema de registros?* Se trata de una pregunta que debe analizarse tanto al considerar la creación de un nuevo registro como al procesar los registros existentes.

5.1 ¿CÓMO SE DEBEN ESTRUCTURAR LAS ENCUESTAS BASADAS EN REGISTROS?

La metodología estadística basada en registros consta de tres niveles:

- *Establecimiento del sistema.* Tiene que ver con la estructuración y coordinación de un gran número de registros en un sistema funcional de registros estadísticos. Todo el sistema debe diseñarse de modo tal que sea posible producir estadísticas nuevas y de mejor calidad.
- *Creación de un registro estadístico dentro del sistema ya existente.* Tiene que ver con la responsabilidad respecto a los métodos aplicados cuando se crea un registro estadístico. ¿Cómo se deben usar los datos administrativos para crear un registro primario? ¿Cómo se debe construir un registro base? ¿Cómo se puede usar un sistema de registros para obtener un registro integrado? El objetivo es generar una o más matrices de datos que constituirán el registro. Veremos este tipo de preguntas en este capítulo y el siguiente.

– *Aplicación de una encuesta basada en un registro estadístico ya disponible.* De acuerdo con los objetivos de investigación, el propósito del trabajo consiste en analizar las matrices de datos del registro a partir de diferentes cálculos, agregaciones o métodos de estimación. ¿Cómo se deben efectuar estos cálculos para tener en cuenta los diferentes problemas metodológicos, por ejemplo, de cobertura, valores perdidos y cambios de nivel en las series de tiempo? ¿Cómo se deben procesar los registros con variables de valores múltiples? Los resultados del procesamiento de un registro arrojan como resultado varias tablas estadísticas para una encuesta específica. Abordaremos estas interrogantes en los capítulos 8 y 9.

Los datos administrativos se deben procesar para satisfacer los requisitos de los propósitos estadísticos. Con base en todas las fuentes disponibles relacionadas con el registro estadístico que se creará, se deben definir de la mejor manera posible el conjunto de objetos, los objetos y el contenido de las variables del registro. ¿Qué procesamiento de registros debe llevarse a cabo? ¿Cómo se debe hacer? En este capítulo y el siguiente, describimos el procesamiento desde una perspectiva estadística y en el capítulo 11 ofrecemos una breve descripción desde una perspectiva informática.

Procedimiento para crear un registro estadístico

¿Cómo se debe crear un registro estadístico que en parte se usará para encuestas específicas y en parte contribuirá a hacer otras encuestas dentro del sistema de registros? Al igual que el trabajo con encuestas por muestreo, la creación de un registro estadístico se divide en distintas etapas:

1. Determinación de los objetivos de investigación: ¿qué requerimientos y objetivos estadísticos se deben cumplir con el registro? (Sección 5.2).
2. Integración del inventario: ¿cuáles son las fuentes disponibles, incluidas las administrativas, cuando se planea la creación de un nuevo registro? ¿Qué registros estadísticos ya presentes en el sistema se pueden usar? (Sección 5.3).
3. Integración de las fuentes administrativas y los registros estadísticos que ya existen en un nuevo registro. Este proceso de integración debe pasar por una fase de planeación que puede subdividirse en tres partes:
   a. Definición del conjunto de objetos del registro. (Sección 5.4).
   b. Definición de los objetos del registro (unidades estadísticas). (Sección 5.5).
   c. Definición del contenido de las variables del registro. (Secciones 6.1 y 6.2).
4. Contactos con los proveedores de datos y recepción de los datos administrativos: verificación y validación de los datos administrativos recibidos. (Sección 6.3).
5. Integración de las distintas fuentes en un nuevo registro estadístico; esta fase se puede subdividir en tres partes:
   a. ¿Cómo se pueden integrar las fuentes que ya existen para que el registro contenga el conjunto de objetos necesario? Los datos administrativos se verifican y validan de modo que se obtenga el conjunto de objetos requerido. Se procede al emparejamiento de las distintas fuentes y se seleccionan los objetos. Se procesan las referencias temporales a fin de crear el conjunto de objetos correspondiente al momento o el periodo de que se trate. (Sección 5.4).
   b. ¿Qué procesamiento debe llevarse a cabo para verificar y corregir las definiciones de los objetos? Los datos administrativos se verifican y validan para que las definiciones de los objetos correspondan a las que se requieren. Entonces se forman los objetos derivados en el nuevo registro. (Sección 5.5).

c. ¿Qué procesamiento debe realizarse para crear las variables en cuestión? Las variables de las fuentes administrativas se verifican y validan. Se toman medidas si faltan valores de las variables. Se emparejan las distintas fuentes, se seleccionan las variables y se importan al nuevo registro. Se forman las variables derivadas en el nuevo registro. (Capítulo 6).

¿Qué procesamiento debe realizarse? El registro que se creará será un registro base, un registro estadístico primario o un registro integrado. Los requisitos de procesamiento pueden variar para estos tres tipos de registro. El trabajo con una encuesta basada en registros se debe estructurar de tal forma que los responsables de la encuesta estén conscientes de los tres procesos que ocurren a la vez. El primer proceso es la creación del registro estadístico, que se describe en los incisos 1 a 5 anteriores. De manera paralela, tanto el registro como el procesamiento deben pasar por una etapa de *aseguramiento de la calidad* y *documentación.*

Aseguramiento de la calidad

Se debe asegurar la calidad del trabajo de creación de registros estadísticos. La calidad del registro se debe revisar y describir usando diversos indicadores. La documentación también representa una parte importante del aseguramiento de la calidad. El uso incorrecto y no discriminado de los datos administrativos se puede evitar mediante los metadatos, que nos dan información sobre posibles problemas de comparabilidad. Abordaremos estos problemas en el capítulo 10.

Documentación

Como muchos usuarios del sistema de registros deben utilizar un registro estadístico, todos los registros se deben documentar para que todos consulten y entiendan la documentación. Los metadatos tienen un papel muy importante en el trabajo con estadísticas basadas en registros. Cuando se vinculan y emparejan diferentes registros, es necesario conocer las definiciones y cualquier problema de comparabilidad que pudiera surgir. También es importante que los métodos de procesamiento se documenten para facilitar la formulación de una metodología y los intercambios de experiencias. En el capítulo 11 veremos lo relativo a metadatos y documentación.

Gráfico 5.1 Los tres procesos paralelos de una encuesta basada en registros

Creación y uso de un registro	Aseguramiento de la calidad (capítulo 10)	Documentación (capítulo 11)
Creación de un registro (capítulos 5 y 6) 1. Determinación de los objetivos de investigación 2. Integración del inventario 3. Planeación 4. Recepción de datos 5. Integración: – crear población – crear objetos – crear variables **Cálculos y tabulaciones (capítulos 7, 8 y 9)** Selección de métodos de estimación para hacer cálculos y compilar tablas	Contactos con proveedores de datos Verificación de datos al recibirlos Causas y magnitud de valores perdidos Causas y magnitud de no emparejamientos Evaluación de la calidad de los objetos y las variables Cuestionarios para el mantenimiento del registro Investigación de inconsistencias entre distintas fuentes Documentación	Al mismo tiempo que se recaban microdatos de los distintos registros fuente, también se deben recabar metadatos para esos registros fuente: – Poblaciones de registro de cada fuente – Definición de las variables importadas – Verificaciones y procesamientos a los que se han sometido las fuentes – Calidad de las fuentes Documentación del registro

5.2 DETERMINACIÓN DE LOS OBJETIVOS DE INVESTIGACIÓN

¿Qué encuestas es necesario hacer? ¿Qué preguntas se deben responder con base en esas encuestas? La definición de los objetivos de investigación debe dar respuesta a estas preguntas. Esta fase es de fundamental importancia para la calidad y la relevancia de la encuesta; sin embargo, muchas veces se pasa por alto. Un error en el que se incurre a menudo es empezar a recabar datos con demasiada anticipación, de modo que la encuesta no permitirá obtener las respuestas necesarias. Antes de invertir en una captación de datos costosa, siempre debemos tratar de usar los datos disponibles. En una encuesta basada en registros podemos trabajar en la definición de los objetivos de investigación y, al mismo tiempo, verificar los datos de interés que ya se conocen.

Una vez concluida la definición de los objetivos de investigación, se habrán determinado tanto la población objetivo como las variables estadísticas importantes. Más adelante en este capítulo analizaremos cómo se define la población objetivo y en el capítulo 6 cómo se definen las variables estadísticas. Estas definiciones deben ser lo más adecuadas y funcionales posible.

Diferentes usuarios necesitan diferentes registros

El gráfico 2.11 del capítulo 2 nos servirá para analizar la manera en que la Oficina de Estadística de Suecia ha llevado a cabo encuestas basadas en registros con objeto de satisfacer las necesidades de diversos usuarios.

En ese gráfico se describen los registros disponibles en la sección del sistema correspondiente a población. Se puede aplicar el mismo razonamiento a otras secciones del sistema y a un registro completamente nuevo. Hasta ahora hemos considerado el gráfico 2.11 como parte de un sistema de registros. No obstante, hay preguntas específicas que han motivado la creación de cada registro. En la Oficina de Estadística de Suecia se acumulan conocimientos sobre las necesidades de los distintos usuarios. Al mismo tiempo, el personal debe conocer el tema, estar familiarizado con las posibilidades del sistema de registros y saber qué metodología estadística podría ser de utilidad. Por consiguiente, el gráfico también muestra la estructura de un área de la encuesta (estadísticas sobre individuos) y sus partes.

1. Registro base:	Estadísticas demográficas: población, nacimientos/defunciones, migración

El registro base establece los cimientos para las estadísticas demográficas, que tienen por objeto describir la estructura de la población y sus cambios con el paso del tiempo. Estas cuestiones demográficas constituyen una categoría aparte de objetivos de investigación con una metodología bien establecida.

2. Registros basados directamente en datos administrativos:	Registro de Ingresos y Tributación Vehículos de propiedad privada Registro de Pacientes Registro de Cáncer Registro de Causas de Defunción

Durante mucho tiempo se han usado varias fuentes administrativas importantes como base para la estadística. La educación y el ingreso son áreas donde se han conformado registros estadísticos sucesivamente en cooperación con diversos investigadores y otros usuarios. Esto también se aplica al uso de datos médicos de registros para fines de investigación.

3. Registros integrados para estadísticas oficiales:	"Censo" basado en registros Registro de Empleo Registro de Educación

Algunos de los funcionarios más creativos de la Oficina de Estadística de Suecia, que mantienen buenos contactos con diferentes usuarios, vislumbran nuevas posibilidades para aprovechar los datos de registros ya disponibles. Cuando logran combinar las nuevas necesidades con las posibilidades del sistema, se crea un nuevo registro integrado.

4. Registros integrados para fines de investigación:	Registro Multigeneracional Registro de Fecundidad Registro Longitudinal de Ingresos Registro Longitudinal de Asistencia Social Transición de los Estudios al Mercado de Trabajo

Los investigadores pueden tener acceso a los registros disponibles en la Oficina de Estadística de Suecia para sus propios análisis. La Oficina también puede generar nuevos registros para satisfacer las necesidades específicas de los investigadores combinando distintos registros. Toda la vinculación entre entradas se lleva a cabo con apego a la Ley de Protección de Datos de Suecia y sólo se divulgan datos anonimizados después de una revisión que garantiza el principio de confidencialidad.

5. Modelos de microsimulación:	El Registro de Ingresos y Tributación constituye la base de un modelo de simulación donde se pueden poner a prueba los cambios planeados en materia tributaria y pagos por concepto de transferencias. Tanto el gobierno como los investigadores académicos usan este modelo. El Registro de Empleo constituye la base de un modelo regional de simulación y pronóstico usado por el gobierno y los investigadores académicos.

Los modelos de simulación basados en datos de los registros de la Oficina de Estadística de Suecia pueden mostrar la manera en que los cambios previstos pueden afectar factores como la distribución del ingreso o el desarrollo en una región específica.

6. Tablas regionales estandarizadas:	El Registro de Población, el Registro de Ingresos y Tributación, los Vehículos de Propiedad Privada, el Registro de Empleo y el Registro de Educación conforman la base para las tablas regionales estandarizadas que usan los gobiernos locales y los analistas del mercado. Las regiones se pueden seleccionar mediante tecnología SIG.

Con diferentes paquetes de tablas se pueden generar estadísticas regionales para cualquier desglose regional. Los perfiles de residencia y de mercado también son ejemplos de productos basados en registros con tablas regionales que los clientes empresariales pueden usar para varios análisis de mercado.

Distintos alcances de los objetivos de investigación

Una encuesta por muestreo tradicional, por ejemplo, la Encuesta sobre Fuerza Laboral, sólo presenta un conjunto de datos, a saber, los datos captados en la encuesta. El equipo responsable de esa encuesta puede describir su misión como *"recabamos datos para la Encuesta sobre Fuerza Laboral, los analizamos y presentamos"*.

Existe el riesgo de que quienes trabajan con encuestas basadas en registros entiendan su misión de forma igualmente estrecha: *"somos responsables de los datos administrativos de la fuente X y presentamos datos de X"*. Si un equipo responsable de una encuesta basada en registros entiende así su misión, no se aprovecharán todas las oportunidades del sistema de registros. Se debe trascender este alcance limitado y usar todos los registros relevantes en el sistema para analizar y describir el tema de que se trate.

Ejemplo: Registro de Estudiantes Universitarios
Este registro capta datos administrativos de los alumnos de todas las universidades sobre qué quieren estudiar, qué estudian realmente y los resultados de sus esfuerzos y exámenes. El enfoque tradicional consiste en sólo presentar esas fuentes; sin embargo, existe más información sobre esta población en el sistema de registros:

1. ¿También trabajan para obtener un ingreso y cuánto ganan?
2. ¿De dónde vienen?
3. ¿Qué hacían antes de entrar a la universidad?
4. ¿Qué ocurrió cuando concluyeron sus estudios universitarios?

Si en vez de sólo presentar información sobre los estudios, el equipo responsable de las estadísticas universitarias también se ocupa de responder estas cuatro preguntas, podrá aprovechar más registros del sistema.

5.3 ELABORACIÓN DE UN INVENTARIO DE LAS FUENTES

Luego de estructurar los objetivos de investigación, se debe elaborar un inventario de las distintas fuentes útiles para crear el nuevo registro. Hay tres tipos de fuentes (registros fuente) que pueden ser de interés:

– Registros estadísticos disponibles en el sistema de registros de una oficina de estadística. La información sobre el contenido de estos registros debe quedar clara en la documentación de los registros disponibles.
– Fuentes administrativas conocidas dentro de la oficina de estadística, pero que en la actualidad no se usan en el sistema de registros. Por ejemplo, puede haber algunos datos entregados que no se usan en este momento en la oficina o partes de los registros administrativos que la oficina no solicita a la autoridad administrativa. Una tercera posibilidad consiste en llevar a cabo una *captación de datos integrada*, es decir, la autoridad administrativa agrega una pregunta a su cuestionario específicamente para fines estadísticos. La información sobre estas posibles fuentes se puede hallar en el área de la oficina de estadística que mantiene contacto con la autoridad correspondiente.
– Registros administrativos completamente nuevos de los que nadie en la oficina de estadística tiene conocimiento actualmente. Pueden ser fuentes administrativas de las empresas o autoridades que aún no se usan para propósitos estadísticos, pero podrían usarse. Estamos hablando de fuentes completamente nuevas, por lo que es importante establecer contactos fuera de la oficina de estadística.

Todas las fuentes disponibles relacionadas con el área problema se deben analizar cuando se crea el nuevo registro. Éste es un principio importante, pues toda fuente nueva podría servir para mejorar la cobertura del nuevo registro en relación con los objetos y las variables. Una nueva fuente también supone la posible detección de inconsistencias, lo que contribuiría a mejorar la calidad.

Una fuente con fallas de calidad podría usarse en combinación con otras fuentes. Tal vez por sí sola la fuente produciría estadísticas de baja calidad, pero *como parte del sistema de registros* podría contribuir a mejorar, en términos generales, la cobertura y la calidad del sistema.

5.4 DEFINICIÓN DEL CONJUNTO DE OBJETOS DE UN REGISTRO

En toda encuesta estadística, se debe definir la población que se encuestará. En esta sección, analizaremos la definición de conjunto de objetos y en la siguiente sección la de objetos o unidades estadísticas.

Si se desea aprovechar un registro para varias encuestas, el registro debe permitir que se definan distintas poblaciones a partir de él, es decir, que sea posible hacer diferentes selecciones de objetos dentro del registro de modo que los diferentes subconjuntos se adecuen a las distintas encuestas. Otra posibilidad consiste en crear varias versiones del registro para diversas aplicaciones.

Reservaremos el concepto *población* para referirnos a un conjunto de objetos que pertenece a una encuesta específica. Cuando describamos un registro sin referirnos a una encuesta específica, usaremos el concepto *conjunto de objetos*.

Sin embargo, cada registro estadístico se crea para uno o varios usos o encuestas principales. Por lo tanto, es común que el conjunto de objetos del registro coincida con la población de la encuesta principal. Cuando se genera un nuevo registro para una encuesta específica, se debe definir la población del nuevo registro. Cada registro fuente tiene su propio conjunto de objetos, que se incluirá total o parcialmente en el nuevo registro.

5.4.1 Definición de una población

Cada encuesta se inicia con una serie de preguntas formuladas de manera teórica o general. A continuación, los conceptos teóricos de las preguntas se deben operacionalizar, es decir, se deben traducir a conceptos medibles. Cuando se definen estos conceptos medibles, se determina *qué* se está encuestando.

> La definición de una población debe mostrar claramente qué objetos están incluidos en esa población. El tipo de objeto también se especificará con claridad. Asimismo, siempre se incluirán una referencia temporal y una delimitación geográfica. Esta última deberá indicar la relación que existe entre los objetos o unidades estadísticas y el área geográfica.

Ejemplo de una definición de población:
"Personas con residencia permanente en Suecia al 31/12/2004. 'Residencia permanente' se refiere a…"

personas	= tipo de objeto
con residencia permanente en	= relación
Suecia	= área geográfica
al 31/12/2004	= momento

La teoría de las encuestas por muestreo y las directrices para los conceptos y declaraciones de calidad de las oficinas de estadística normalmente incluyen tres conceptos relativos a las poblaciones:

- *Población de interés*: Población a la que se refiere la pregunta teórica de que se trate.
- *Población objetivo*: Población operacionalizada, o población teórica de interés, traducida a una población concreta que se examinará, es decir, la población *objetivo* de la encuesta.
- *Población marco*: Conjunto de objetos al que realmente da origen el *marco*. Estos conceptos se pueden hallar en la teoría de las encuestas por muestreo y censos con una captación de datos propia.

En el caso de las encuestas basadas en registros, sólo usamos los dos primeros conceptos. Como en las encuestas basadas en registros no hay marco muestral, el concepto *población marco* se debe remplazar por:

– *Población de registro*: Conjunto de objetos del registro creado para la encuesta de que se trate, es decir, la población que *realmente* se está encuestando.

Hay diferencias importantes entre una población marco y una población de registro. Una población marco se define *antes* de la captación de datos, mientras que la población de registro se crea *después* del periodo de referencia, cuando se han recibido todos los datos administrativos. Un marco puede incluir objetos distintos de los que se encuestarán, por ejemplo, un mapa o un registro de domicilios para una encuesta sobre hogares.

Ejemplo: Estructura de la población usada en el Registro de Población
En este caso, queremos hacer una encuesta sobre el tamaño y la estructura de la población sueca al 31 de diciembre de un año específico. La *población de interés* son los residentes permanentes en Suecia al 31 de diciembre, pero se trata de un concepto vago que se debe precisar más. Esto se puede hacer de distintas maneras. En general, una buena definición debe satisfacer los siguientes requisitos:

– ser *adecuada*, es decir, concordar con aquello que se desea encuestar;

– ser *funcional*, es decir, ser aplicable en un sentido práctico.

Cuando se elabora una definición, muchas veces es difícil encontrar el equilibrio entre lo que queremos encuestar (definición adecuada) y lo que es posible encuestar (definición funcional). En este ejemplo, serían posibles las siguientes definiciones de *residentes permanentes en Suecia al 31/12/año t*:

(1) Personas registradas por el Consejo de Tributación de Suecia al 31/12/año *t*

(2) Personas registradas por el Consejo de Tributación de Suecia al 31/12/año *t*, de acuerdo con los datos disponibles a finales de enero del año *t+1*

El Registro de Población sueco suele esperar hasta el final de enero para crear la versión correspondiente a la situación al 31/12. Se espera que para entonces se hayan notificado todos los cambios y eventos que afectan el registro del año anterior. El registro generado debe ser la versión que se aplique aunque todavía se pueden recibir algunas notificaciones referidas al año *t*. Por lo tanto, a esta encuesta se aplica la definición (2) de *población de registro*.

No obstante, en la declaración sobre calidad para las estadísticas demográficas anuales suecas se señala que la encuesta tiene por objeto describir la *población objetivo* de acuerdo con la definición (1) anterior. Por ende, la diferencia entre población objetivo y población de registro constituye el *error de cobertura* del registro. Las defunciones y emigrantes no notificados antes de finalizar enero causan una *sobrecobertura*, mientras que los nacimientos e inmigrantes no notificados originan una *subcobertura*.

Además de estos errores de cobertura, puede presentarse un error de relevancia si no se aplica la definición adecuada. La diferencia entre población de interés y población objetivo es uno de los *errores de relevancia* de la encuesta. Probablemente haya entre 25 000 y 50,000 personas incluidas en el Registro de Población de Suecia que no residen permanentemente en el país. Se calcula que de 4 a 8% de inmigrantes procedentes de países no nórdicos han dejado Suecia sin notificarlo. Este error de relevancia afecta las estimaciones estadísticas sobre tasas de defunción, ingreso promedio, etc. para los inmigrantes originarios de países no nórdicos, de modo que las estimaciones se vuelven engañosas.

En el ejemplo anterior, observamos que las estadísticas poblacionales de Suecia usan una definición administrativa —la población registrada— cuando definen la población objetivo de la encuesta. Los conceptos administrativos siempre dan definiciones funcionales. Se considera una práctica estadística sensata usar estos conceptos administrativos para definir la población objetivo si los errores de relevancia son pequeños. Sin embargo, la regla básica es que al definir la población se debe tratar de satisfacer las demandas de la encuesta estadística. Si los conceptos administrativos no son lo suficientemente relevantes o adecuados, será necesario formular definiciones propias y procesar el registro de tal manera que el conjunto de objetos refleje con el mayor apego posible la población definida.

¿Es posible mantener la definición administrativa actual de la población sueca? Éste es un tema de análisis en la Oficina de Estadística de Suecia; hay maneras de mejorar la definición de la población objetivo:

– Incluir estudiantes extranjeros que asisten a universidades suecas, pues hay datos administrativos disponibles. Son estudiantes no registrados en el Consejo de Tributación, pero sí en las universidades.
– Excluir a los estudiantes suecos en el extranjero, que están registrados en el Consejo de Tributación, pero no residen de manera permanente en Suecia. Hay datos administrativos disponibles para muchos de ellos.
– En muchos casos, los estudiantes universitarios suecos pueden estar registrados en el Consejo de Tributación con el domicilio donde vivían antes de entrar a la universidad, mientras que en la universidad tienen registrado su domicilio actual. Esto reubicaría geográficamente a muchos estudiantes universitarios y, por consiguiente, modificaría las estadísticas de población regionales.

Lo anterior significa que es posible mantener la definición administrativa de la población objetivo e introducir una nueva definición, que será más adecuada.

5.4.2 ¿Pueden alterarse los datos del Consejo Nacional Tributario?

En caso de ser responsables de un registro en una oficina de estadística, ¿podemos modificar los datos administrativos procedentes de otra autoridad? Las autoridades administrativas recaban los datos y, por lo tanto, les cabe la responsabilidad original del registro. Un funcionario de una oficina de estadística que trabaja en un producto o encuesta específica podría pensar que no puede modificar o complementar los datos con objeto de adecuarlos a los propósitos estadísticos del producto en cuestión.

Sin embargo, en el capítulo 1 señalamos que los registros administrativos se deben procesar para que los objetos y las variables correspondan a las necesidades de la estadística. Esto significa que el personal de una oficina de estadística que recibe datos administrativos tiene tanto la libertad como la obligación de efectuar esas modificaciones a fin de mejorar la calidad de las estadísticas. Claro está, las personas que hacen estos cambios deben ser experimentadas e independientes y contar con el apoyo de una red de estadísticos especializados en registros. De lo contrario, quizá no se atrevan a hacer los cambios.

Puede resultar muy cómodo no cambiar en absoluto los datos administrativos que se reciben y después responsabilizar de todos los errores a la autoridad administrativa. Por otro lado, si modificamos los datos que recibimos, podemos sentirnos responsables de todos los errores. No obstante, si somos estadísticos, siempre seremos responsables de las estadísticas que producimos.

5.4.3 Definición de una población: registros primarios

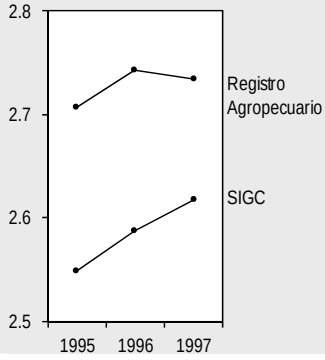
¿Cómo deben definir población objetivo y población de registro quienes trabajan con registros primarios? Los registros primarios se basan total o parcialmente en registros administrativos, por lo que existe el riesgo de que los conjuntos de objetos del sistema administrativo influyan de forma inapropiada en la selección de la población de registro.

Quizás el conjunto de objetos en el registro administrativo no cubra por completo la población objetivo de interés estadístico. El conjunto de objetos administrativo consiste exclusivamente en aquellos objetos incluidos en el sistema administrativo y, en consecuencia, puede haber tanto una sobrecobertura como una subcobertura en comparación con la población objetivo estadísticamente deseable.

Ejemplo: ¿El conjunto de objetos administrativo es adecuado como población objetivo?
Las estadísticas agrícolas se basan en las solicitudes de subsidios que los agricultores de la Unión Europea presentan a los consejos administrativos locales. Estas solicitudes se registran ante el Sistema Integrado de Gestión y Control (SIGC), mediante el cual se administran los subsidios agrícolas.

En el gráfico 5.2 se comparan datos incluidos en las solicitudes de subsidios por área en el periodo 1995-1997 dentro del SIGC y los datos correspondientes del Registro Agropecuario, basado en un censo de la Oficina de Estadística de Suecia. Las diferencias entre el SIGC y el Registro Agropecuario obedecen a la subcobertura en el registro del SIGC, es decir, al hecho de que algunos agricultores no solicitan subsidios aunque están activos.

Gráfico 5.2 Subcobertura en un registro administrativo



Superficie cultivada de tierras de labranza
(millones de hectáreas)

	SIGC	Registro Agropecuario	Cobertura
1995	2.548	2.707	94.1%
1996	2.587	2.743	94.3%
1997	2.618	2.734	95.8%

Cambio porcentual entre años

	SIGC	Registro Agropecuario
95/96	1.5%	1.3%
96/97	1.2%	–0.3%

Conclusiones: Aunque se puede considerar que el registro del SIGC tiene una buena cobertura, variaciones aparentemente pequeñas en la cobertura hacen que la serie de tiempo para la superficie cultivada de tierras de labranza sea totalmente equivocada, pues una disminución real en 1996–1997 parece ser un aumento continuo en el registro del SIGC.

No se deben pasar por alto las fallas de cobertura en los registros administrativos. Por el contrario, en estos casos las poblaciones objetivo se deben definir de acuerdo con los requisitos estadísticos. Luego se debe crear un nuevo registro que contenga la población de registro deseada usando un registro base actualizado, en el caso que estamos analizando seleccionando objetos del Registro Empresarial. Entonces este nuevo registro se puede emparejar con el registro SIGC, lo que permitirá detectar tanto una sobrecobertura como una subcobertura en el SIGC.

La sobrecobertura es un indicio de posibles fallas en el Registro Empresarial. La subcobertura en el registro del SIGC aparecerá en forma de valores perdidos en el nuevo registro.

Esta no respuesta se puede corregir ya sea haciendo un ajuste por no respuesta o agregando una encuesta especial para recabar datos sobre la parte de la población objetivo no incluida en el registro administrativo.

5.4.4 Definición de una población: registros integrados

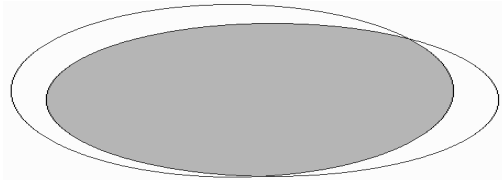
¿Cómo se deben definir población objetivo y población de registro en el caso de los registros integrados? A continuación analizamos los principios básicos aplicables.

Ejemplo: Registros especiales con emparejamiento y selección de una población objetivo
Muchos registros especiales suponen el procesamiento combinado de varios registros. Se lleva a cabo una serie de emparejamientos y se importan variables de diferentes registros. El resultado final es un registro integrado con muchas variables de interés para los clientes del proyecto.

Pero ¿cómo se ha definido la población objetivo de la encuesta? No es difícil que el conjunto de objetos del registro integrado sea una intersección de los conjuntos de objetos de los registros emparejados. ¿Representa esta intersección una población objetivo equivalente para el proyecto? No se debe dar por hecho que así sea.

Un registro especial de este tipo también se debe iniciar definiendo la población objetivo de acuerdo con el problema que se estudiará. Después se seleccionará un conjunto de objetos apropiado del registro base relevante. Este conjunto de objetos será la población de registro, que se emparejará con los registros que contienen las variables de interés. En el caso de objetos que arrojan coincidencias, los valores de las variables se importarán al nuevo registro. En el caso de objetos que no arrojan coincidencias, se evidenciará la no respuesta parcial, es decir, los valores de las variables estarán perdidos.

Gráfico 5.3 Conjuntos de objetos al emparejar dos registros



Cada elipse representa un conjunto de objetos de uno de dos registros distintos.

El área sombreada representa los objetos que arrojan coincidencias cuando se emparejan, es decir, la intersección.

Conclusiones: ¿Qué población objetivo corresponde a la intersección? La intersección no puede ser nunca una población objetivo, porque entonces sería imposible definir esa población. Tampoco debe ser la población de registro, pues se correría el riesgo de pasar por alto la no respuesta a causa de no emparejamientos. Es importante estar conscientes de la no respuesta y también realizar un ajuste por no respuesta cuando se conoce el alcance y la estructura de éstas. Abordaremos este tema en el capítulo 8.

5.4.5 Uso necesario de los registros base para definir poblaciones

Tanto en el trabajo con registros integrados como en la generación de registros primarios a partir de registros administrativos se deben usar los registros base al definir los conjuntos de objetos. Esto obedece a dos razones: en primer lugar, a que los conjuntos de objetos del registro base deben ser los “mejores” y, en segundo, a que las estadísticas basadas en registros deben ser consistentes. Quienes trabajan con registros primarios y registros integrados deben empezar con cada conjunto de objetos del registro base y usar una de las *poblaciones estandarizadas* que se crean para uso general:

– la *versión de fin de año* adecuada para las estadísticas de inventario anuales (como la población al 31/12);
– la *versión por año calendario* adecuada para las estadísticas de flujos anuales (como el ingreso de la población en un año específico);
– una *versión mensual o trimestral* adecuada para estadísticas mensuales o trimestrales.

Metodología general:

1. Definir la población objetivo.
2. Seleccionar el conjunto de objetos deseado del registro base para establecer la población de registro.
3. Hacer un emparejamiento con los registros que contienen variables interesantes.
4. Cuando se identifican coincidencias: importar los valores de las variables al registro que se está generando.
5. Cuando se identifican no emparejamientos: mostrar los valores perdidos (no respuesta parcial).

5.4.6 Requisitos para un registro base

Un registro base debe cumplir con cuatro requisitos para su uso en la definición de poblaciones de registro:

– Contener referencias temporales, es decir, se debe establecer la fecha de ocurrencia de todos los eventos que afectan a los objetos o las unidades estadísticas del registro;
– Tener una buena cobertura;
– Tener variables de vinculación de buena calidad;
– Tener variables de expansión de buena calidad.

Referencias temporales

Se debe conocer el momento en que cada objeto nació o dejó de existir. También se debe conocer el momento en el que ocurrieron otros eventos importantes, como movimientos o categorías modificadas.

La fecha de los eventos es importante, aunque no siempre es posible determinarla. La fecha en que se registró un cambio reviste casi la misma importancia. Un registro base debe contener varios tipos de referencias temporales: fecha del evento, fecha de registro y fecha de cambio. A partir de estas referencias temporales, es posible generar poblaciones de registro que reflejen el estado de la población en un momento o periodo específicos.

Las referencias temporales pueden tener diferentes niveles de precisión. Por lo general, se conocen los eventos de las personas, de modo que se puede precisar un día; en cambio, quizá sólo se sepa el mes en el que ocurrieron los eventos de las empresas.

En un registro base es importante diferenciar entre cambios estadísticamente relevantes y estadísticamente triviales. El cambio de un código postal o una clave telefónica nacional para toda un área no debe aparecer como si el objeto se hubiera movido. Del mismo modo, el cambio de régimen jurídico de una empresa no debe tomarse como la desaparición de una empresa y la fundación de una empresa nueva.

Cobertura, variables de expansión y variables identificadoras

La mayoría de las poblaciones para las estadísticas de la Oficina de Estadística de Suecia se definen usando uno de los registros base. Por ello, es un requisito fundamental que los conjuntos de objetos de los registros base tengan una buena cobertura de las necesidades de muchas encuestas y que los vínculos y variables usados para seleccionar las subpoblaciones o dividir las poblaciones en dominios de interés (*variables de expansión*) estén actualizados y sean de buena calidad. De lo contrario, puede haber errores de cobertura en las subpoblaciones. Esto ocurrirá, por ejemplo, si faltan datos sobre la rama industrial o si la clasificación industrial es incorrecta.

El ejemplo de la sección 5.4.3 sobre empresas agrícolas ilustra los requisitos que debe cumplir un registro base. En este caso, el Registro Empresarial deberá poder usarse para seleccionar empresas agrícolas y generar un nuevo registro con una buena cobertura de la población objetivo. Para asegurar lo anterior, es necesario que el Registro Empresarial tenga una buena cobertura de todas las empresas activas; además, la clasificación industrial debe ser de buena calidad para permitirnos seleccionar la población de registro deseada. Por otro lado, el Registro Empresarial incluye variables identificadoras como número de organización, nombre, domicilio y teléfono de la empresa, todas muy importantes cuando se efectúa el emparejamiento con el registro del SIGC, que contiene solicitudes de subsidios agrícolas.

A fin de asegurar una buena cobertura en un registro base, se deben utilizar todas las fuentes administrativas relevantes. El siguiente es un principio general para crear un registro estadístico: *la calidad puede mejorarse combinando múltiples fuentes.*

Ejemplo: Se considera que la versión de noviembre del Registro Empresarial sueco es de buena calidad, por lo que esa versión se usa como marco de muestreo para diversas encuestas. Esta buena calidad no necesariamente se aplica a las empresas pequeñas y todos los sectores industriales. Por ejemplo, si la versión de noviembre del año *t* se empareja con el registro del SIGC del verano del año *t*, desde luego habrá una subcobertura significativa. Si se suman los datos del Registro de Estados de Resultados (RER), el Registro del Impuesto al Valor Agregado (IVA) y las Declaraciones de Ingresos para Empresas (DIE) para el año *t*, mejorará la cobertura del Registro Empresarial.

Gráfico 5.4 Resultados de diferentes emparejamientos entre el Registro Empresarial y el registro del SIGC

Registro del SIGC emparejado con:	Variables de vinculación	Coincidencias en el SIGC
Registro Empresarial, noviembre 1995	Número de organización	75.2%
Registro Empresarial complementado con:		
RER, IVA y DIE para 1995	Número de organización	96.6%

Conclusiones: Una de las conclusiones a las que podemos llegar a partir de este ejemplo es que el Registro Empresarial para 1995 tiene una cobertura deficiente de las empresas agrícolas. Al complementar el Registro Empresarial con otras fuentes, la cobertura mejora considerablemente. En este ejemplo hay cuatro fuentes que podrían usarse para optimizar esta parte del Registro Empresarial. El registro del SIGC es la fuente más actualizada disponible alrededor de un año antes que las otras fuentes. Además, las empresas incluidas en el SIGC son empresas agrícolas activas, de manera que el registro de ese sistema también contiene información sobre la clasificación industrial.

Si alguna de las variables de un registro base —utilizadas a menudo para los procesos de selección o emparejamiento— están desactualizadas o incompletas, se deben aplicar *cuestionarios de actualización del registro*. Estos cuestionarios, que deben ser breves para no

convertirse en una carga innecesaria para el informante, pueden enviarse a los objetos del registro base al que le faltan datos o está desactualizado.

Ejemplo: Para mantener la calidad del Registro Empresarial, cada año se envían dos cuestionarios a todas las empresas con más de un establecimiento. Aunque estos cuestionarios tienen el propósito de mantener la calidad de la población de establecimientos, también se pregunta a las empresas sobre los detalles actuales en la industria y el domicilio.

En el caso del Registro Empresarial, es natural que, en sus encuestas para empresas, la Oficina de Estadística de Suecia dedique más tiempo a mantener los datos de registro sobre empresas grandes con una importancia económica considerable, pero en una economía dinámica que cambia constantemente este enfoque puede dar pie a errores. Johansson (2001) nos da el ejemplo siguiente, en el que se estudia el crecimiento del sector de tecnologías de la información.

Ejemplo: Los pequeños pueden ser importantes
Johansson describe el crecimiento del sector de tecnologías de la información entre 1993 y 1998. La tabla siguiente se basa en datos del Registro Empresarial (Johansson, pp. 82–83).

Gráfico 5.5 Empleo dentro de la industria de las tecnologías de la información por tamaño de la empresa

Tamaño según número de empleados	0	1	2–4	5–9	10–19	20–49	50–99	100–199	200–499	500+	Total
1993: Número de empleados y trabajadores por cuenta propia	9,157	2,219	6,724	7,548	8,082	10,678	10,123	6,544	16,597	82,588	160,230
1998: Número de empleados y trabajadores por cuenta propia	17,825	3,084	8,459	9,634	11,582	16,434	13,206	12,933	21,045	85,600	199,802
Cambio 1993–1998 (número)	8,668	865	1,735	2,086	3,500	5,756	3,083	6,389	4,478	3,012	39,572
Cambio 1993–1998 (%)	95%	39%	26%	28%	43%	54%	30%	98%	27%	4%	25%

Conclusión: El gráfico 5.5 ilustra que las empresas más pequeñas registran el mayor crecimiento en términos del número de empresas. Esto significa que puede ser riesgoso no considerar a las empresas pequeñas cuando se actualiza un registro. Las mediciones de crecimiento, que resultan importantes para el análisis económico, serían de baja calidad. El ejemplo anterior presenta varios problemas metodológicos importantes: ¿cómo hacer comparaciones entre años?, ¿cómo medir el empleo y el crecimiento? La comparabilidad es un gran problema en el estudio realizado por Johansson (2001), pues el Registro Empresarial presenta cambios de nivel en las series de tiempo correspondientes al periodo 1996-1997, que afectan el estudio de las empresas pequeñas.

Ejemplo: Población del Registro Empresarial

Número de empresas (según el Registro Empresarial)		
	Serie antigua	Serie nueva
1995	562,765	
1996	585,571	
1997	*601,385*	791,385
1998		810,337
1999		797,338

La población de registro del Registro Empresarial consiste en empresas que están obligadas a pagar el IVA, son empleadores o pagan el impuesto para empresas. Desde 1996, toda empresa está obligada a pagar el IVA independientemente de su volumen de facturación; anteriormente había un límite de facturación anual de 200,000 coronas.

Al aplicar este nuevo principio —es decir, cuando se incorporaron al registro las nuevas empresas obligadas a pagar el IVA—, se sumaron alrededor de 190,000 empresas en 1997 para obtener un total de 791,385. Por consiguiente, las reglas administrativas respecto al IVA han definido la población de registro. De acuerdo con el ejemplo del gráfico 5.4, en la antigua serie de empresas agrícolas había una subcobertura de aproximadamente 25 por ciento.

Mediante la combinación de varias fuentes, podría reducirse la subcobertura y tal vez usar la definición de población relevante antes y después de los cambios en el pago obligatorio del IVA.

La cobertura puede variar durante el año

Un registro base se apoya en fuentes que pueden referirse a periodos completamente distintos y presentar diferentes desfases respecto a la publicación de resultados. De modo que la cobertura del registro variará durante el año.

Ejemplo: Cobertura variable en el Registro Empresarial

El Registro Empresarial se basa en los datos mensuales de las declaraciones del IVA de los empleadores, los gravámenes aplicables a éstos y las deducciones fiscales de los empleados. Sin embargo, en el caso de las empresas pequeñas, que no presentan declaraciones de impuestos mensuales, se usan las declaraciones anuales presentadas al año siguiente de que percibieron sus ingresos.

El Registro Empresarial recibe información continua sobre empresas nuevas y reestructuradas con desfases diferentes. Estas nuevas empresas y la reestructuración de empresas ya existentes es algo que ocurre particularmente al final del año calendario. La introducción de estos cambios requiere tiempo y esfuerzo.

Todos los registros base deben cumplir con los requisitos que analizamos en esta sección, aunque el ejemplo al que más recurrimos sea el del Registro Empresarial.

5.4.7 Todos deben apoyar los registros base

Todo el sistema debe "cooperar" en las definiciones de conjuntos de objetos y objetos. Las encuestas por muestreo y los censos toman sus poblaciones marco de los registros base y las encuestas basadas en registros hacen lo propio con sus poblaciones de registro. Quienes realizan encuestas sobre estas poblaciones adquieren conocimientos que deben transmitir al equipo responsable del registro base, de modo que los registros base sean las fuentes óptimas de conjuntos de objetos dentro de la oficina de estadística.

La siguiente es una lista de todos los flujos de información que contribuyen a que el registro base sea de buena calidad en lo relativo a los objetos y los conjuntos de objetos. Quienes trabajan con registros base deben tratar de obtener la mayor cantidad de información posible y quienes trabajan con otros registros del sistema deben aportar información que mejore la calidad del registro base en cuanto a los conjuntos de objetos.

1. Dentro de cada registro base, se debe mantener contacto con las *autoridades* que entregan los datos administrativos. Los cambios en el sistema administrativo pueden afectar los objetos y la cobertura del registro.
2. Dentro de cada registro base, se puede obtener información de *registros base adyacentes*, donde los objetos relacionados pueden dar información sobre los objetos del primer registro.
   Ejemplos: Los cambios en la propiedad de viviendas residenciales también implican cambios en las variables correspondientes del Registro de Población. Si en el Registro de Actividades están perdidas determinadas actividades correspondientes a una persona, esto puede indicar una sobrecobertura de esa persona en el Registro de Población.
3. Los resultados que se obtienen de la verificación de los diferentes *registros estadísticos* puede dar información sobre errores de cobertura o definiciones de objetos incorrectas.

4. El correo devuelto y las razones de "no contacto" son maneras en que las *encuestas por muestreo* y los *censos* pueden dar información relativa a sobrecobertura e indicar si el objeto ha sufrido algún cambio.
5. Los *cuestionarios para el mantenimiento de registros* se realizan mediante un registro base en ciertas partes del conjunto de objetos donde se sospecha que los datos son incompletos o están desactualizados.

Por lo tanto, el registro base desempeña un papel fundamental en relación con las definiciones y los conjuntos de objetos, tema que abordaremos en la siguiente sección. Para asegurar que los registros base sean de la máxima calidad posible, todos los usuarios deben enviar información sobre cualquier falla de calidad que identifiquen. Como veremos ahora, hay tres categorías de usuarios que deben enviar esta información.

Información de encuestas por muestreo y censos

Quienes hacen encuestas por muestreo y censos tienen contacto directo con los proveedores de datos. La correspondencia devuelta y el rastreo en busca de entrevistas telefónicas puede dar información acerca de sobrecobertura, domicilios desactualizados, reestructuración de un objeto u otros cambios de interés para los responsables del mantenimiento del registro base correspondiente. La definición de hogar que se usa en los registros se puede comparar con las encuestas por muestreo dirigidas a los hogares. Las unidades empresariales incluidas en los registros se pueden cotejar con las diferentes encuestas por muestreo para empresas.

Ejemplo: Sobrecobertura en el Registro de Población

Se analizó la no respuesta en las encuestas sobre fuerza laboral y se obtuvo una estimación del número de personas nacidas en el extranjero que representaban una posible sobrecobertura en el Registro de Población. También se analizó el correo devuelto y la no respuesta en un cuestionario para personas nacidas en el extranjero y se obtuvieron estimaciones de la sobrecobertura por país de nacimiento, edad y sexo.

Información de registros primarios

Los encargados de los registros primarios reciben los registros administrativos, que pueden contener objetos no presentes en el registro base. Además, pueden haber desaparecido objetos del registro administrativo, lo que quizás indique que el objeto ha cambiado, se ha reestructurado o dejó de existir.

Ejemplo: Sobrecobertura en el Registro de Población

Se analizó más a fondo a quienes no tienen ingreso disponible de acuerdo con el Registro de Ingresos y se obtuvo una estimación de sobrecobertura para distintas categorías de personas nacidas en el extranjero. Se demostró la importancia de corregir la sobrecobertura para que las estadísticas sobre ingresos por país de nacimiento no resulten engañosas.

Información de registros integrados

Quienes trabajan con registros integrados pueden comparar valores de variables relacionados con el mismo objeto, procedentes de distintas fuentes. Los errores e inconsistencias detectados pueden obedecer a cambios en el objeto.

Ejemplo: Definiciones de objeto incorrectas en el Registro Agropecuario

El registro del SIGC contiene solicitudes de empresas agrícolas para recibir subsidios. Wallgren (1999) emparejó este registro con el Registro Agropecuario de la Oficina de Estadística de Suecia utilizando como variables de vinculación el NIP y el número de la organización. Se verificaron todas las coincidencias comparando el área de tierra cultivable en ambos registros. Se identificaron alrededor de 9% de falsas coincidencias. Al combinar

unidades con el mismo número telefónico, se pudieron derivar nuevas unidades para las que la superficie de tierras de labranza era correcta. En la sección 5.5 abordaremos las definiciones de objetos.

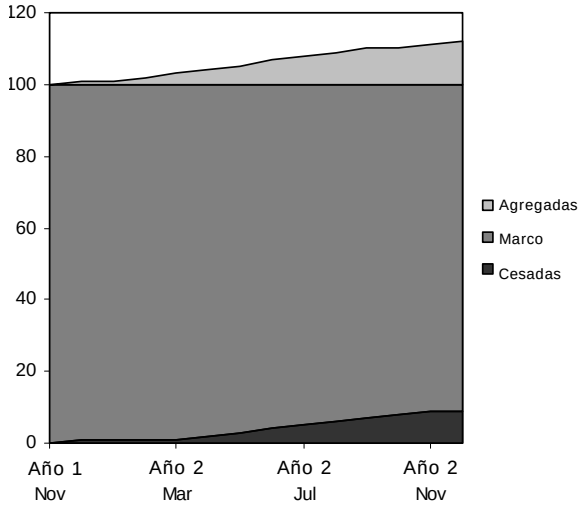
5.4.8 Problemas de cobertura en encuestas con captación de datos propia

En las encuestas con captación de datos propia, la población marco se define *antes* de captar los datos. Normalmente, las encuestas empresariales trimestrales y mensuales se basan en la versión del inventario actual del Registro Empresarial disponible en noviembre. Las encuestas deben referirse a un mes o un trimestre durante determinado año, al que llamaremos año 2. Por consiguiente, la población objetivo estará formada por empresas activas durante el periodo de referencia en el año 2. Para estas encuestas, la población marco se define como el inventario real de empresas activas, de acuerdo con la versión del Registro Empresarial de noviembre del año 1.

Gráfico 5.6 Sobrecobertura y subcobertura en encuestas con captación de datos propia

Cambios en la población de registro entre noviembre del año 1 y diciembre del año 2

Porcentaje de la población marco



Las empresas agregadas después de noviembre del año 1 generan una *subcobertura* (gris claro).

Parte *restante* de la población marco (gris intermedio).

Las empresas del marco que dejaron de existir después de noviembre del año 1 generan una *sobrecobertura* (gris oscuro).

En el ejemplo del gráfico 5.7 se observan los problemas de sobrecobertura y subcobertura que presentan las encuestas con una captación de datos propia. La población marco que se usará se decide *antes* de captar los datos y se compara con la población de registro que se crea *después* del periodo de la encuesta, cuando ya se ha recibido toda la información administrativa. La población marco que se define en noviembre del año 1 se usa en encuestas anuales con una captación de datos propia para el año 1, así como en encuestas de ese mismo tipo para el trimestre 1 del año 2.

El Registro Empresarial recibe la información sobre restructuración, apertura y cierre de empresas con un desfase bastante amplio. Además, los datos sobre la clasificación industrial de la empresa también pueden ser incorrectos en el marco de noviembre, lo cual se detectará durante el año 2 o el 3. En cierto momento en el otoño posterior al año en cuestión, se habrá recibido casi toda la información sobre el año anterior y será posible crear un registro por año calendario de buena calidad.

Si la población marco de la tabla A se usa para el año 1, el número de identificación (NI) 1 será sobrecobertura, el NI 5 será subcobertura y la rama industrial para NI 3 será incorrecta; en cambio, si se usa para el año 2, los NI 1 y 2 serán sobrecobertura, los NI 5 y 6 serán subcobertura y la rama industrial para NI 3 será incorrecta.

Gráfico 5.7 Poblaciones marco y registros anuales

A. Población marco establecida en noviembre del año1 para los años 1 y 2

NI empresarial	Rama industrial
NI 1	DE
NI 2	DB
NI 3	***DA***
NI 4	DC
-	-
-	-

B. Registro por año calendario generado en el otoño del año 2 respecto al año 1

NI empresarial	Rama industrial
-	-
NI 2	DB
NI 3	***DB***
NI 4	DC
NI 5	DG
-	-

C. Registro por año calendario generado en el otoño del año 3 respecto al año 2

NI empresarial	Rama industrial
-	-
-	-
NI 3	***DB***
NI 4	DC
NI 5	DG
NI 6	DC

5.4.9 Problemas de cobertura en las encuestas basadas en registros

Si comparamos las encuestas con captación de datos propia y las encuestas basadas en registros, la principal diferencia estriba en que la población de registro de estas últimas se crea *después* del periodo de referencia relevante. Dependiendo de la rapidez con la que el sistema administrativo recibe la información sobre el surgimiento de nuevos objetos y la desaparición de antiguos objetos, al cabo de un tiempo es posible crear una población de registro relacionada con un periodo o momento específicos. Los problemas de cobertura derivados de que la población marco se defina *antes* del periodo de referencia —como ocurre siempre en una encuesta con captación de datos propia— no existen en las encuestas basadas en registros. Sin embargo, las fallas en el sistema administrativo pueden hacer que se pierdan ciertas categorías de objetos o que no se notifiquen cambios importantes para ciertos objetos.

Gráfico 5.8 Definiciones de población en diferentes tipos de encuesta

	Ventajas	Desventajas
Estadísticas basadas en encuestas con captación de datos propia	Pueden actualizarse.	Se presentan problemas considerables de sobrecobertura y subcobertura, así como errores en las variables de expansión si los cambios se registran de manera tardía.
Estadísticas basadas en registros	Tienen buena cobertura y variables de expansión más precisas.	En ciertos casos, hay un gran desfase entre la ocurrencia del evento y el momento en que las estadísticas están disponibles.

Un registro de población, creado correctamente, siempre tendrá mayor calidad que la población marco correspondiente porque se basa en información mejor y más abundante.

5.5 DEFINICIÓN Y DERIVACIÓN DE OBJETOS

Unidad estadística y objeto estadístico son sinónimos, aunque en este libro usamos el término *objeto*. Los tipos de objetos del sistema de registros pueden crearse de distintas maneras:

1. En un sistema administrativo con propósitos administrativos. Estos objetos suelen tener una relevancia jurídica.
2. En cooperación con una autoridad administrativa para propósitos estadísticos.

3. Dentro de una oficina de estadística mediante la captación de información que ha permitido definir los objetos.
4. Dentro de una oficina de estadística mediante el procesamiento del registro. Esta categoría recibe el nombre de *objetos derivados*.

5.5.1 Tipos de objetos administrativos

Muchos de los tipos de objetos administrativos son de interés estadístico, como también lo pueden ser las variables de estos registros administrativos.

Tipos de objetos administrativos de interés estadístico:
Personas, empleos (trabajadores asalariados o por cuenta propia), propiedades o vehículos.

Tipos de objetos administrativos no siempre de interés estadístico:
El concepto de una familia en el marco de la ley tributaria (pareja casada o que cohabita y tiene hijos en común) no corresponde a un hogar de interés estadístico. El tipo de objeto *unidad jurídica* en un registro empresarial administrativo quizá tampoco sea estadísticamente relevante, pues una empresa puede presentar sus declaraciones de impuestos al Consejo Tributario usando más de una identidad de unidad jurídica.

5.5.2 Tipos de objetos creados en cooperación con autoridades administrativas

Los tipos de objetos estadísticos creados por una oficina de estadística en cooperación con una autoridad administrativa tienen la ventaja de que ciertos datos administrativos se pueden vincular con estos objetos de interés estadístico, de modo que es posible describir estos tipos de objetos con variables estadísticas sin necesidad de una captación de datos que implique un uso intensivo de recursos.

Tipos de objetos estadísticos creados en cooperación con las autoridades administrativas:
La Oficina de Estadística de Suecia crea establecimientos a partir de la información recabada por el Consejo Nacional Tributario. Esta información se capta por medio de los estados de resultados anuales de todos los empleadores. Éste es un ejemplo de captación de datos integrada, es decir, hay una colaboración entre la oficina nacional de estadística y una autoridad administrativa.

5.5.3 Tipos de objetos estadísticos creados mediante la captación de datos propios

Cuando se usan encuestas con una captación de datos propia, es posible captar datos sobre tipos de objetos con un interés estadístico. Para describir estos tipos de objetos con variables estadísticas, es necesaria una captación de datos con uso intensivo de recursos.

Tipos de objetos estadísticos creados mediante captación de datos propia:
Los hogares y las viviendas son dos clases de tipos de objetos creados en las encuestas en hogares. Dentro de las estadísticas económicas, las unidades empresariales, las unidades de tipo de actividad y las unidades de tipo de actividad local se forman luego de entablar contacto con las empresas.

Todos los registros administrativos empresariales se relacionan con objetos llamados unidades jurídicas, que se identifican por el número de organización. Para propósitos estadísticos, los tipos de objeto *unidades empresariales*, *establecimientos*, *clase de actividad* y *clase de actividad local* se forman dentro del Registro Empresarial. Todas estas unidades tienen su propio número de identidad y deben ser observables para poder captar datos. Se requiere un trabajo intensivo para mantener un alto nivel de calidad.

5.5.4 Tipos de objetos estadísticos derivados de la oficina de estadística

Los tipos de objetos estadísticos creados por derivación dentro de la oficina de estadística se basan en objetos administrativos. Los datos administrativos para los objetos administrativos se pueden usar entonces para describir los objetos estadísticos derivados.

Hogares derivados

Son los hogares identificados en el registro de población de personas por vivienda como se hace actualmente en Dinamarca y Finlandia. Si hay datos administrativos en los que todas las personas de una población están registradas en una vivienda específica, es posible definir hogares derivados como aquellos registrados en la misma vivienda. Esta posibilidad es crucial para los "censos" basados en registros. La Oficina de Estadística de Dinamarca (Statistics Denmark, 1995) y la Oficina de Estadística de Finlandia (Statistics Finland, 2004) describen cómo generar información censal tradicional mediante encuestas basadas en registros en caso de que sea posible generar hogares derivados.

Unidades empresariales derivadas

Si queremos usar fuentes administrativas para estadísticas económicas, es muy importante crear unidades empresariales derivadas. Todas las fuentes administrativas proporcionan datos sobre unidades jurídicas, como propiedad y relaciones entre ellas, que deben usarse para derivar unidades empresariales. Una unidad empresarial puede constar de una o más de estas unidades jurídicas.

En el siguiente gráfico vemos un ejemplo de la manera en que los datos administrativos de una empresa se notifican al Consejo Tributario mediante distintas unidades jurídicas. Para comparar la facturación y la suma salarial, es necesario sumar los valores de todas las unidades jurídicas pertenecientes a esa unidad empresarial.

Gráfico 5.9 Unidad empresarial constituida por siete unidades jurídicas

		Facturación (miles de millones de coronas)		Suma salarial (miles de millones de coronas)	
		Fuente 1	Fuente 2	Fuente 1	Fuente 3
Unidad empresarial 1	Unidad jurídica 1				0.1
Unidad empresarial 1	Unidad jurídica 2	8.6		1.3	
Unidad empresarial 1	Unidad jurídica 3		0.2		
Unidad empresarial 1	Unidad jurídica 4		1.8		0.2
Unidad empresarial 1	Unidad jurídica 5		4.2		0.6
Unidad empresarial 1	Unidad jurídica 6		1.7		0.3
Unidad empresarial 1	Unidad jurídica 7		0.9		0.1
	Suma:	**8.6**	**8.8**	**1.3**	**1.3**

5.5.5 Objetos e identidades: requisitos para un registro base

Además de la importancia que tiene un registro base dentro del sistema en lo que respecta al conjunto de objetos, de acuerdo con la sección 5.4.6 también se crean objetos derivados dentro de los registros base. Otra tarea relevante y difícil consiste en dar seguimiento a los objetos y sus identidades a lo largo del tiempo, así como hacer un registro estadístico de cambios relevantes.

Todos los tipos de objetos deben tener una definición clara y precisa. Cuando se usan tipos de objetos administrativos para fines estadísticos, es necesario familiarizarse con las definiciones del sistema administrativo. Los responsables de crear objetos dentro de una oficina

de estadística también se ocupan de definir estos tipos de objetos. Se deben documentar tanto las definiciones administrativas como las estadísticas.

Al emparejar registros, es importante que una *misma identidad* presente en los diferentes registros realmente *se relacione con el mismo objeto*. Si el emparejamiento arroja coincidencias falsas y se importan valores de variables de un objeto a otro con una identidad similar, los resultados de la encuesta serán incorrectos. El riesgo de que ocurran estas falsas coincidencias es particularmente elevado cuando se emparejan registros de diferentes momentos o periodos, en los que puede haber cambios en las identidades.

Aunque en distinto grado, este tipo de problema se presenta en todas las partes de un sistema de registros. En casos raros, el mismo número de identificación personal en los registros de distintos años puede referirse a diferentes personas. También pueden ocurrir cambios en el número de identificación personal, de modo que la misma persona tenga distintos números en las diferentes versiones anuales de un registro. Una tarea importante para los responsables de un registro base consiste en mantener un registro de referencias cruzadas con los números de identidad antiguos y nuevos. El Registro de Población sueco contiene estos datos de referencias cruzadas para números de identificación personal, que se pueden utilizar cuando se emparejan registros sobre individuos.

En el Registro Empresarial se presentan los mismos problemas: una misma empresa puede operar con diferentes números de identidad a la vez y estos números pueden cambiar con el tiempo. Por ello, se debe mantener un registro de referencias cruzadas con diferentes tipos de números de identificación y dar seguimiento a los cambios a lo largo del tiempo.

Las identidades inmobiliarias pueden constar de condado, municipio y parroquia[1]. Estos códigos de propiedad cambiarán en caso de que se modifiquen las divisiones regionales, lo cual significa que cuando se emparejan registros de diferentes años, se deben usar claves entre los códigos regionales anteriores y los nuevos.

El Registro de Actividades se ve afectado por los cambios tanto en los números de identificación personal como en las identidades de las empresas. Cuando este registro se empareja con el Registro Empresarial, debe ser posible usar las identidades de establecimientos de ambos para vincular establecimientos idénticos.

5.5.6 Objetos e identidades: requisitos para los registros primarios e integrados

Cuando se reciben los registros administrativos para crear registros primarios, se deben seguir estos pasos a fin de que los *objetos* del registro estadístico se apeguen a las definiciones especificadas:

- Se verifica que el formato de los números de identificación de los registros administrativos sea correcto. Se deben validar los números duplicados.
- Se empareja el registro administrativo con el registro base que contiene el mismo tipo de objetos. Este emparejamiento se hace contra la versión del registro base correspondiente al mismo periodo. Se verifican las identidades de los objetos que producen un no emparejamiento y se corrigen usando, por ejemplo, nombre, domicilio y teléfono.
- Si el registro estadístico debe relacionarse con un tipo de objeto diferente del que aparece en la fuente administrativa, puede resultar apropiado unir o dividir los valores para los objetos administrativos. Esto se hace por medio del registro base, donde los objetos administrativos se vinculan con el tipo de objeto de que se trate.

[1] Nota de la traductora: subdivisión de los municipios que en el pasado coincidía territorialmente con las parroquias de la Iglesia sueca. Actualmente, las parroquias se utilizan como demarcaciones para fines estadísticos.

Ejemplo: En el gráfico 5.10 tenemos datos (sumas salariales) que describen unidades administrativas (unidades jurídicas, UJ). A fin de describir las sumas salariales de las unidades estadísticas (empresas, UE), se agregan las sumas salariales de todas las unidades jurídicas pertenecientes a la misma unidad empresarial.

Gráfico 5.10 Agregación de datos correspondientes a las unidades jurídicas para describir unidades empresariales

Registro base	
Id. de UE	**Id. de UJ**
UE1	UJ11
UE1	UJ12
UE2	UJ21
UE2	UJ22

Registro administrativo	
Id. de UJ	**Sumas salariales**
UJ11	180
UJ12	450
UJ21	270
UJ22	500

Registro estadístico	
Id. de UE	**Sumas salariales**
UE1	180 + 450 = 630
UE2	270 + 500 = 770

Nota: Las diversas variables económicas están adaptadas en mayor o menor grado para la agregación. Si, por ejemplo, se agregan nuevos pedidos, los pedidos a un subcontratista se contarán dos veces.

Ejemplo: El gráfico 5.11 muestra que es posible usar un modelo para dividir la facturación de una unidad jurídica (UJ) por establecimiento (Es). En este caso, se usa como modelo el número de empleados y la facturación de la rama industrial por empleado para empresas de cierta rama. La elección del modelo es subjetiva: podría haberse usado la suma salarial en vez del número de empleados.

Gráfico 5.11 División de los datos para unidades jurídicas entre establecimientos

Registro Base				Modelo		Registro Administrativo		Registro Estadístico			
Id. de UJ	Id. de establecimiento	Rama industrial	Núm. de empleados	Rama industrial	Facturación/empleado	Id. de UJ	Facturación	Id. de establecimiento	Rama industrial	Empleados	Facturación calculada a partir de un modelo
UJ1	Es11	DG	120	DA	2.1	UJ1	300	Es11	DG	120	$\frac{300\cdot120\cdot1.5}{(120\cdot1.5+60\cdot2.1)} = 176.5$
UJ1	Es12	DA	60	DB	1.2	UJ2	250	Es12	DA	60	$\frac{300\cdot60\cdot2.1}{(120\cdot1.5+60\cdot2.1)} = 123.5$
UJ2	Es21	DD	50	DD	1.8			Es21	DD	50	$\frac{250\cdot50\cdot1.8}{(50\cdot1.8+100\cdot1.2)} = 107.1$
UJ2	Es22	DB	100	DG	1.5			Es22	DB	100	$\frac{250\cdot100\cdot1.2}{(50\cdot1.8+100\cdot1.2)} = 142.9$

Los usuarios de datos de varios registros del sistema, ya sea para crear un registro integrado o para importar datos a su "propio" registro, deben verificar diversos aspectos a fin de detectar errores e inconsistencias respecto a los conjuntos y definiciones de los objetos:

- Se debe verificar que los conjuntos de objetos de los diferentes registros se refieran al mismo momento o periodo.
- Se deben investigar los objetos que no arrojan una coincidencia: ¿por qué no la hubo? Si todos los registros se crearon con el registro base respectivo, no debe haber ningún no emparejamiento, aunque puede seguir habiendo valores perdidos debido a un no emparejamiento anterior con el registro base.
- Se deben verificar los valores de los diferentes registros, respecto a variables relacionadas o similares, para identificar coincidencias falsas.
- Se deben notificar los errores e inconsistencias identificados a los responsables de los registros de que se trate.

5.5.7 Cómo hacer un emparejamiento: aspectos generales

Hemos estado hablando del *emparejamiento exacto*, cuyo propósito es encontrar objetos idénticos en los diferentes registros, o bien objetos que tienen una relación definida. También existe el *emparejamiento estadístico*, que tiene por objeto hallar objetos *similares* para fines analíticos. Cuando creamos registros dentro de un sistema de registros estadísticos, el método de emparejamiento debe ser exacto.

Al crear un registro estadístico, se usa la vinculación de datos o el emparejamiento (exacto) para dos propósitos distintos:

– Combinar distintas fuentes a fin de crear un conjunto de objetos —es decir, la población de registro— con una buena cobertura. Esto lo hemos abordado en este capítulo.
– Usar distintas fuentes para crear las variables del nuevo registro. Veremos esto en el capítulo 6.

El proceso de emparejamiento se debe planear con cuidado: qué variables de vinculación se usarán y en qué orden se combinarán las distintas fuentes. La calidad de las variables de vinculación es importante, al igual que su validación.

Hay que investigar las causas y la extensión del no emparejamiento, y decidir si las unidades no emparejadas se deben incluir o no en la población de registro. Si se incluyen, el no emparejamiento dará como resultado unidades con valores perdidos para algunas variables. También se deben verificar los objetos aparentemente emparejados, pues de lo contrario las falsas coincidencias generarán errores.

Ejemplo: Emparejamiento de diferentes registros agropecuarios
Wallgren (1999) analizó los problemas de emparejamiento relacionados con las solicitudes de subsidios registradas en el SIGC, el Registro Agropecuario y el Registro Empresarial.

Las variables identificadoras se deben validar antes del emparejamiento. Antes de validar los números telefónicos, sólo 47% de los agricultores del Registro Agropecuario pudieron emparejarse con las unidades correspondientes del registro del SIGC. Luego de validar los números telefónicos, fue posible emparejar a 64 por ciento.

Al combinar dos variables identificadoras (número telefónico con la granja y el número de identidad fiscal de las granjas), mejoró el resultado del emparejamiento, de modo que 96% de las unidades del registro del SIGC pudieron emparejarse con unidades del Registro Agropecuario. Esto significa que mejorando los métodos de emparejamiento fue posible aumentar los emparejamientos exitosos al principio de 47% a 64% y después hasta 96 por ciento.

También se deben verificar los vínculos. Un emparejamiento entre variables identificadoras no es prueba suficiente de que los objetos del SIGC y el Registro Agropecuario son idénticos. Si el objeto del SIGC tiene una superficie de tierras cultivadas mayor que la del objeto del Registro Agropecuario, tal vez el objeto del SIGC deba vincularse con dos objetos del Registro Agropecuario y viceversa. Los vínculos se verificaron comparando la superficie total de tierras cultivadas, la superficie de tierras cultivadas confiable y la ubicación descrita por parroquia.

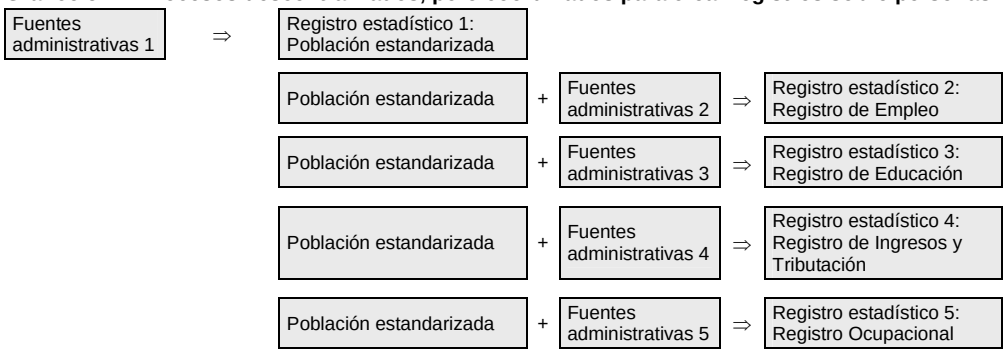
5.6 CÓMO PRODUCIR ESTADÍSTICAS REGIONALES BASADAS EN REGISTROS

La Oficina de Estadística de Suecia produce estadísticas regionales mediante un proceso de estimación descentralizado, lo que da como resultado microdatos completamente consistentes y coherentes. Los principios presentados en este capítulo se aplican al trabajo con diferentes registros que incluyen datos sobre individuos. Gracias a los métodos usados, es posible generar estadísticas regionales detalladas y de total consistencia.

En la primera etapa del proceso, el equipo responsable del Registro de Población crea una población estandarizada, que servirá de base para quienes trabajan con los otros registros.

La población estandarizada se define como la población al 31 de diciembre —la población al 31 de diciembre del año *t* se crea a principios de febrero del año *t+1*— y se usa como población de registro en otros registros estadísticos, como vemos en el gráfico 5.12.

Gráfico 5.12 Procesos descentralizados, pero coordinados para crear registros sobre personas



Como las variables estadísticas de las fuentes administrativas 2–5 no se traslapan, los registros estadísticos 2–5 se pueden crear de manera independiente. Los cinco registros del gráfico 5.12 se pueden integrar en un solo registro con todas las variables.

Gráfico 5.13 Registro integrado con partes de diferentes registros sobre personas

NIP	Variables de clasificación del Registro de Población					Variables estadísticas											
						Registro de Empleo			Registro de Educación			Registro de Ingresos y Tributación			Registro Ocupacional		
	var1	var2	var3	var…	var…	var…	var…	var…	var…	var…	var…	var…	var…	var…	var…	var…	var…
1																	
2																	
3																	
…																	
N																	

Las variables de clasificación regional del Registro de Población se crearon uniendo códigos y coordenadas regionales del Registro Inmobiliario para cada individuo del Registro de Población. El hecho de que las personas estén registradas en el inmueble en el que viven facilita el proceso de emparejamiento. El Consejo Nacional Tributario es responsable de este registro; como el lugar donde las personas están registradas es importante desde el punto de vista administrativo, los códigos y coordenadas regionales son de buena calidad.

Se produce un gran número de tablas estandarizadas para satisfacer las demandas de distintos tipos de usuarios. La división regional es flexible porque se pueden usar coordenadas. Como veremos en el capítulo 12, se debe controlar el riesgo de revelar datos confidenciales.

CAPÍTULO 6

Cómo crear un registro: las variables

De acuerdo con la sección 5.1, la creación de un registro estadístico se divide en cinco fases. En la fase 5c, se crea el contenido de las variables. ¿Qué procesamiento es necesario para generar esas variables? Las variables de las fuentes administrativas se deben verificar y validar. Posteriormente, se emparejan las distintas fuentes, se seleccionan las variables y se importan al nuevo registro, y, por último, se crean las variables derivadas en el nuevo registro.

6.1 DETERMINACIÓN DEL CONTENIDO DE LAS VARIABLES DEL REGISTRO

Una variable es una característica medible de un objeto. En la sección 3.3, se describe el concepto de *variable* desde el punto de vista de las ciencias estadísticas y la informática. En el sistema de registros, distinguimos entre variables dependiendo de su origen:

– *variables primarias locales:* tomadas de un registro administrativo o de un censo para incluirlas en el registro en cuestión;
– *variables derivadas localmente:* formadas en el registro en cuestión usando otras variables del registro;
– *variables primarias importadas:* variables primarias tomadas de otro registro del sistema y llevadas al registro en cuestión;
– *variables derivadas importadas:* variables derivadas tomadas de otro registro del sistema y llevadas al registro en cuestión.

6.1.1 Definición de variables

Toda encuesta comienza por una serie de preguntas formuladas en términos teóricos o generales. A continuación deben operacionalizarse los conceptos teóricos de las preguntas, es decir, traducirse a conceptos medibles. Cuando se definen estos conceptos medibles, se determina *qué* se está encuestando. Ya nos habíamos referido a esto cuando explicamos cómo se define la población de una encuesta. En el caso de la definición de variables, también es necesario que los conceptos teóricos relacionados con las características de los objetos se traduzcan a variables estadísticas claramente definidas.

Al definir una variable se debe especificar con qué tipo de objeto se relaciona y su momento o periodo de referencia, es decir, la característica debe relacionarse con un momento o periodo específico. Además, se debe detallar cómo medir la característica y con qué escala. Veamos dos ejemplos relacionados con la edad para ilustrar lo anterior:

Definición 1: La edad se refiere a la edad de una persona en años completos cumplidos al final del año en que se realiza la encuesta. *Ejemplo:* Una persona nacida el 1 de enero de 1980 tiene, para efectos de una encuesta referida al 31/12 de 2000, 20 años cumplidos.

Definición 2: La edad se refiere a la edad de una persona redondeada a una posición decimal en el momento de referencia. *Ejemplo:* Una persona nacida el 1 de enero de 1980 tiene, para efectos de una encuesta referida al 31/12 de 2000, la edad de 21.0 años.

La *definición de una variable primaria* se determina cuando se formula una pregunta en un cuestionario o un formulario administrativo. Las instrucciones para la pregunta también se incluyen en la definición. Por lo tanto, cuando se documentan estadísticas con base en datos administrativos, los cuestionarios y folletos con las instrucciones de las autoridades administrativas siempre se deben recopilar y guardar, de preferencia también de manera electrónica.

La *definición de una variable derivada* está constituida en parte por las definiciones de las variables en las que se basa y en parte por la regla con la que se formó.

Para todos los tipos de variables también es deseable que las definiciones sean *adecuadas*, es decir, que concuerden con lo que se encuestará, y *funcionales*, es decir, que sean fáciles de usar.

Definiciones adecuadas de variables

Si la definición de una variable es adecuada, esto significa que *concuerda con otras variables* del registro creadas para efectos de la encuesta. En un registro empresarial, las definiciones de variables como producción, ventas e inventario de entrada y de salida deben concordar de modo que las relaciones lógicas entre las variables sean aplicables. La producción menos ventas debe ser igual al cambio registrado en el inventario para el periodo.

Como todo registro estadístico forma parte de un sistema, *las definiciones de las variables en los diferentes registros también deben concordar* para ser adecuadas. Por ejemplo, en todos los registros sobre individuos es una regla básica que variables como edad, estado civil, etc. se definan de la misma manera. De lo contrario, se dificultaría el uso combinado de los datos de registro y disminuiría la calidad de las estadísticas basadas en registros por una falta de consistencia y comparabilidad.

Cuando los valores de las variables cualitativas se agrupan o éstas se dividen en intervalos de clase, los grupos o intervalos de clase deben ser iguales en los diferentes registros. Por ejemplo, si las categorías de edad son las mismas en los diferentes registros sobre individuos, las tablas publicadas serán comparables y, en consecuencia, aumentará la coherencia de las estadísticas, que es un aspecto importante del concepto de calidad.

Definiciones de variables funcionales y variables derivadas

Siempre resulta funcional usar las definiciones de las variables administrativas. Si éstas no satisfacen completamente las necesidades estadísticas, se debe tratar de formar variables derivadas (véase sección 3.3.3). Para crear variables de interés estadístico, se deben explorar las posibilidades que ofrece el sistema de registros. Importando variables de los registros fuente tanto administrativos como estadísticos, se puede crear un registro rico en contenido y generar muchas variables derivadas de interés.

Ejemplo: El Registro de Ingresos y Tributación (descrito en la sección 1.4.1) se basa en cientos de variables de ingresos administrativos tributarios. A partir de ellas se pueden formar diversas variables estadísticas, como ingreso disponible.

Denominación de variables y variables estandarizadas

Las variables de un registro usadas para diferentes encuestas o productos deben estar *estandarizadas*. Esto significa que la responsabilidad por la denominación, la calidad y la documentación de la variable le corresponde al registro donde la variable se creó dentro de la oficina de estadística. Por consiguiente, para ese registro la variable estandarizada será una variable primaria local o una variable derivada localmente.

Cuando otras unidades de la oficina de estadística usan una variable estandarizada, no deben modificar ni su nombre ni su definición. Todos los funcionarios de la oficina de estadística deben tener fácil acceso a la documentación. Quienes importan una variable estandarizada no tendrán necesidad de producir su propia documentación para esa variable. El nombre de la variable estandarizada no deberá usarse como nombre de otras variables.

6.1.2 Creación de las variables de un registro

Una vez creada y validada la población del nuevo registro, se procede a la selección de variables. Las variables de interés estadístico se importan de diversas fuentes.

1. Los objetivos de investigación determinan cuál será el contenido del registro. ¿Qué necesidades tienen los usuarios? ¿Qué posibilidades hay de proyectos futuros?
2. Las variables para el nuevo registro se deben tomar de todas las fuentes relevantes. Cuando se crea un registro estadístico, se deben usar todos los datos de registro que resulten relevantes para los objetivos de investigación. Se crearán las mejores variables posibles. El registro estadístico deberá limpiarse de todo error obvio y ser consistente, es decir, no contener datos contradictorios. También se deben importar otras variables que se usarán para la validación. Estas variables para validación se correlacionan con las variables del registro, de modo que se pueden formular reglas de validación. Las variables de versiones anteriores del registro en cuestión se deben usar para fines de validación.
3. Las variables derivadas adjuntas se forman mediante emparejamiento con otros registros, mientras que las variables derivadas agregadas se forman emparejando y procesando otros registros. Estos tipos de variables derivadas se describen en la sección 3.3.3.
4. Las variables del registro se validan. ¿Cómo detectar errores obvios y sospechados? Veremos lo relativo a la validación en la sección 6.3.
5. ¿Cómo remplazar valores ilógicos o perdidos? En la sección 8.1 describimos cómo abordar la no respuesta parcial en los registros.
6. Las variables derivadas se deben formar usando las variables del registro, como veremos en la sección 6.2.
7. Se deben documentar el procesamiento del registro, los resultados de la validación y las variables del nuevo registro.

Los registros creados serán parte del sistema de registros, lo que implica que deberán cumplirse ciertos requisitos relacionados con la coordinación y la cooperación. La validación y el procesamiento deben ser adecuados para otros usos en el sistema de registros. Esto ayuda a evitar la duplicación del trabajo y aumenta el valor del registro que se crea.

6.2 FORMACIÓN DE VARIABLES DERIVADAS A PARTIR DE MODELOS

En la sección 3.3.3 hablamos de las variables derivadas y describimos cuatro maneras de crearlas:

1. Derivación de variables por agrupamiento de valores o división en intervalos de clase.
2. Derivación de una variable (estadística) por medio de cálculos y procedimientos lógicos con diversas variables (administrativas) de la matriz de datos. En esta sección analizaremos este tipo de variables derivadas.
3. Derivación por adjunción de una variable de otro registro referido a otro tipo de objetos. Por ejemplo, la rama industrial correspondiente a una persona asalariada se establece especificando la rama industrial del establecimiento donde la persona está empleada. Una variable adjunta puede ser de valores múltiples.
4. Derivación por agregación de una variable en otro registro. Por ejemplo, el ingreso del hogar se establece sumando los valores del ingreso personal de todos los individuos que constituyen un hogar.

Mediante una serie de variables administrativas, es posible formar una variable estadísticamente significativa. En esta sección usamos y para la variable estadística derivada y x_1, x_2,… para las variables administrativas. Distinguimos entre situaciones donde y es una variable cualitativa o cuantitativa, y entre valores de variables calculados con exactitud o estimados a partir de un modelo.

La derivación de valores de variables por medio de cálculos tiene que ver con la *imputación* de valores a las variables. La diferencia radica en que la variable derivada se crea para *todos* los objetos de un registro, mientras que el valor imputado de una variable sólo se genera para los objetos del registro donde hay valores perdidos. En la sección 6.2.4 comparamos la derivación y la imputación. Las secciones siguientes también son de interés para el cálculo de valores imputados.

6.2.1 Cálculo exacto de valores para una variable obtenida por medio de una regla

En muchas situaciones resulta apropiado formar una nueva variable aplicando una regla, que debe estar bien fundamentada y documentada.

Ejemplo, variable cualitativa:
En el Registro de Población sueco, la variable *nacidos en el extranjero* o *nacidos en el país*, y, se crea a partir de otras cuatro variables:

Gráfico 6.1 Clasificación de nacidos en el extranjero y nacidos en Suecia

y		x_1	x_2	x_3	x_4
Nacido en el extranjero o en el país		**País de nacimiento**	**Tiempo de residencia**	**País de nacimiento del padre**	**País de nacimiento de la madre**
Código:					
1.1	Nacido en el extranjero con residencia de 0–4 años	Nacido en el extranjero	0–4 años	-	-
1.2	Nacido en el extranjero con más de 5 años de residencia	Nacido en el extranjero	5 años o más	-	-
2.1	Nacido en el país con ambos padres nacidos en el extranjero	Nacido en el país	-	Nacido en el extranjero	Nacida en el extranjero
2.2	Nacido en el país con un padre nacido en el país y otro nacido en el extranjero	Nacido en el país	-	Nacido en el extranjero Nacido en el país	Nacida en el país Nacida en el extranjero
2.3	Nacido en el país con ambos padres nacidos en el país	Nacido en el país	-	Nacido en el país	Nacida en el país

Ejemplo de variable cuantitativa:
El ingreso disponible *y* se calcula aplicando una regla para mostrar qué variables de ingreso x_{i1}, x_{i2},… se deben sumar y qué impuestos x_{t1}, x_{t2}… se deben restar.

En los dos ejemplos anteriores, se formulan reglas sobre la manera en que se definirá la variable derivada. Estas reglas se basan en el conocimiento de la materia y el criterio. Los cálculos son precisos, es decir, si las variables *x* no presentan ningún error de medición, lo mismo debe ocurrir con la variable *y*.

6.2.2 Estimación del valor para una variable derivada mediante una regla

Las reglas para definir una variable derivada también se deben aplicar en los dos ejemplos siguientes. Las reglas utilizadas son modelos completamente basados en el conocimiento de la materia. No obstante, los cálculos no son precisos; por ejemplo, aunque las variables *x* no presenten errores de medición, la variable *y* puede tenerlos. Estos errores en la variable *y* reciben el nombre de *errores del modelo*. En la sección 6.2.3 abordaremos los modelos basados en el análisis causal.

Ejemplo: Ocupación en el sector público, dando prioridad a las fuentes
Los registros de personal de los empleadores del sector público contienen variables administrativas usadas por la Oficina de Estadística de Suecia para clasificar a los empleados por ocupación de acuerdo con la Clasificación Internacional Uniforme de Ocupaciones (CIUO). Las variables administrativas *nombre del puesto* y *código TNS* se han usado de acuerdo con ciertas reglas, modificadas en el año 2000. A continuación presentamos una comparación de las reglas anteriores y las nuevas, cuando se establecen prioridades en las variables administrativas de diferentes maneras:

Gráfico 6.2 Clasificación ocupacional en el sector público

Hasta 2000	**Desde 2001**
Primera etapa: (Alrededor de 90% de los casos.) Existen y se usan nombres de puesto aplicables.	Primera etapa: (Alrededor de 48% de los casos.) Existe y se usa el código TNS*.
Ejemplo: El nombre del puesto: 93460 “editor de sitios web” se vuelve CIUO: 2451 “autores, periodistas y otros escritores”	*Ejemplo:* TNS: 1316 “diseñar, probar y documentar programas” se vuelve CIUO: 3121 “técnico en programación informática”
Segunda etapa: (Alrededor de 8% de los casos). No hay nombre de puesto aplicable, pero existe y se usa el código TNS.	Segunda etapa: (Alrededor de 51% de los casos). No hay código TNS, pero existe y se usa el nombre de puesto aplicable.
Ejemplo: TNS: 1316 “diseñar, probar y documentar programas” se vuelve CIUO: 3121 “técnico en programación informática”	*Ejemplo:* Nombre del puesto: 93460 “editor de sitios web” se vuelve CIUO: 2451 “autores, periodistas y otros escritores”
Tercera etapa: (Alrededor de 2% de los casos). No existe un nombre de puesto útil ni código TNS. Se vuelve no respuesta, CIUO perdida.	Tercera etapa: (Alrededor de 1% de los casos). No existe código TNS ni nombre de puesto aplicable. Se vuelve no respuesta, CIUO perdida.

* Nota de la traductora: código ocupacional utilizado por los empleadores suecos.

Aunque el código TNS y el nombre de puesto sean correctos, en ocasiones la ocupación de acuerdo con la CIUO puede ser incorrecta. No hay una relación exacta entre dos variables administrativas y la ocupación real de un individuo.

Así pues, las reglas no se pueden dar por hecho; las variables administrativas pueden usarse de acuerdo con diferentes principios y puede haber cambios de nivel en las series de tiempo. Aunque la CIUO siga siendo la misma, puede haber cambios de nivel en las series de tiempo cuando se perfeccionan las reglas. Si mejora la calidad del código TNS, conviene dar prioridad a esta variable.

De la misma manera, hay reglas para convertir los códigos ocupacionales de municipios, consejos de condados y empleadores privados al sistema CIUO.

Ejemplo: Situación de empleo en el Registro de Empleo 1985–1992

En la primera versión del Registro de Empleo (1985), la intención era usar los datos de los estados de resultados para medir el empleo como en un censo tradicional: al menos una hora de empleo remunerado durante la semana de medición en noviembre. Los datos de los estados de resultados que incluían información sobre todas las operaciones entre el empleador y los empleados se interpretaron de la siguiente manera:

- Estado de resultados para una parte del año incluido noviembre: si el ingreso mensual total (considerando el subsidio por enfermedad) superaba las 200 coronas, el individuo se clasificaba como empleado remunerado en noviembre.
- Estado de resultados para todo el año: si el ingreso anual total era superior a las 21 800 coronas, el individuo se clasificaba como empleado remunerado en noviembre.
- Si no se cumplía ninguna de las condiciones anteriores, el individuo se clasificaba como no empleado remunerado en noviembre.
- Había también reglas especiales para marinos y trabajadores por cuenta propia.

Aunque todas las variables del estado de resultados fueran correctas, la regla anterior podía inducir a clasificaciones incorrectas. Las personas que tienen un empleo remunerado de acuerdo con la definición del Censo de Población y Vivienda podían ser clasificadas incorrectamente como empleados no remunerados y viceversa. Sin embargo, se hizo una evaluación y se consideró que la calidad de la variable derivada para empleo era aceptable. Entre 1986 y 1992, se aplicaron estas reglas, pero con límites para el ingreso ajustados de acuerdo con un índice salarial.

La variable estadística *situación de empleo* se basa en tres variables administrativas: momento al que corresponden los datos incluidos en el estado de resultados (primer y último mes del periodo con el que se relaciona el empleo), salario bruto y subsidio por enfermedad. La regla establece cómo se deben interpretar estas variables para clasificar a un individuo como empleado en un trabajo remunerado o no empleado en un trabajo remunerado en noviembre.

6.2.3 Estimación del valor de una variable derivada con un modelo causal

En los dos ejemplos de la sección 6.2.2, se formulan reglas para definir una variable derivada. Estas reglas se basan en el conocimiento de la materia y el criterio individual. Una alternativa para el uso de estas reglas consiste en analizar la relación entre la variable deseada y y las variables administrativas (que en este caso llamamos x_1, x_2,…) mediante el diseño de un modelo (causal) estadístico. A partir del modelo estadístico, se puede crear entonces la variable derivada.

Cuando se crean variables derivadas por medio de un modelo estadístico, hay dos etapas en las que se usan diferentes matrices de datos:

1. La primera matriz con *datos de prueba* de, por ejemplo, una encuesta por muestreo que contiene tanto la variable *y* como las variables *x*. Con esta matriz, primero se arma un modelo con el propósito de mostrar la manera óptima de estimar *y* para los valores dados de las variables *x*.
2. Ese modelo se usa en la segunda matriz, la *matriz de datos del registro*, donde sólo existen las variables *x*. Con el modelo estimado, se calcula un valor *y* para cada objeto del registro, a partir de los valores *x* conocidos del objeto.

La ventaja de un modelo estadístico respecto a una regla, basada en el conocimiento de la materia y el criterio, es que un buen modelo estadístico indica la mejor manera de usar muchas variables administrativas. El modelo puede contener muchas variables, a diferencia de una regla basada enteramente en el conocimiento de la materia. La desventaja radica en que es necesario generalizar los resultados de análisis obtenidos con los datos de prueba para aplicarlos a la matriz de datos del registro, es decir, el hecho de que el modelo sea bueno para los datos de prueba no significa necesariamente que será bueno para la matriz de datos del registro.

Variables derivadas cuantitativas

Cuando *y* es una variable cuantitativa, hay tres tipos diferentes de modelos que pueden usarse:

– *Valores medios relacionados con un grupo*: La relación entre la variable *y* y las variables *x* se estudia en una tabla que incluye los valores medios de *y* con base en los datos de prueba.

– *Cocientes:* Si es posible suponer que *y* es proporcional a una variable *x*, se podrá usar el cociente $\Sigma y/\Sigma x$ estimado de los datos de prueba. Se pueden calcular diferentes cocientes para diferentes grupos de objetos.

– *Modelo de regresión:* Si es posible construir un modelo de regresión $\hat{y} = f(x_1, x_2, \ldots)$ con los datos de prueba, se puede usar la función estimada $f(x_1, x_2, \ldots)$.

Ejemplo: Consumo de energía en la industria, cocientes relacionados por grupo

Se lleva a cabo una encuesta por muestreo entre establecimientos con 10 a 49 empleados. Entonces es posible contar con información sobre el número de empleados, la rama industrial y el consumo de energía para la muestra, que consiste en alrededor de 1 800 establecimientos. Entre los establecimientos con *menos de 10 empleados* (alrededor de 50 000 establecimientos), no se selecciona una muestra, sino que se da por sentado que, dentro de cada rama industrial, se consume la misma cantidad de energía por empleado que en las empresas que tienen de 10 a 49 empleados.

En este caso, una muestra de 1 800 empresas representa los datos de prueba que se usan con objeto de calcular los cocientes $\Sigma y/\Sigma x$ = consumo de energía/número de empleados para diferentes ramas de la industria. Estos cocientes se usan para calcular el consumo de energía estimado (=coeficiente estimado · número de empleados en el establecimiento) para cada establecimiento con menos de 10 empleados. Una desventaja de esto es que un modelo que funcione para empresas con 10 a 49 empleados no necesariamente será adecuado para empresas muy pequeñas.

Variables derivadas cualitativas

El ejemplo anterior del Registro de Empleo hace que nos preguntemos si aplicando la regla realmente se usan de manera óptima las variables administrativas. Si, por ejemplo, la información relativa al tiempo en los datos sobre estados de resultados es de mala calidad, ¿debe haber una regla que dependa mucho de esta variable? Al analizar los datos de prueba de una encuesta por muestreo, es posible estudiar la relación entre el empleo remunerado y las variables de los estados de resultados.

Si *y* es una variable cualitativa, se pueden aplicar los siguientes tipos de modelos:

– *Modelo de regresión:* si *y* sólo tiene dos categorías, la relación se puede estudiar usando un modelo de regresión $\hat{y} = f(x_1, x_2,\ldots)$.

– *Análisis discriminante:* si *y* tiene más de dos categorías, se puede recurrir a un análisis discriminante para estudiar la mejor manera de usar las variables *x* para clasificar los objetos en las diferentes categorías *y*.

– *Modelo de minería de datos:* también se pueden usar programas de cómputo para la minería de datos.

Ejemplo: Situación de empleo en el Registro de Empleo de 1993

Como vimos anteriormente, en la primera versión del Registro de Empleo se aplicaron las siguientes reglas:

– Estado de resultados para una parte del año incluido noviembre: si el ingreso mensual total (considerando el subsidio por enfermedad) superaba las 200 coronas, el individuo se clasificaba como empleado remunerado en noviembre.

– Estado de resultados para todo el año: si el ingreso anual total era superior a las 21 800 coronas, el individuo se clasificaba como empleado remunerado en noviembre.

Sin embargo, esta metodología basada en reglas tiene grandes desventajas. Para empezar las personas jóvenes y de edad avanzada se clasifican de acuerdo con los mismos límites de ingresos pese a que sus niveles salariales son distintos. Las pautas de los ingresos, la distribución entre empleos permanentes y temporales y la estructura de los datos de los estados de resultados han cambiado con el tiempo, al igual que el sistema de subsidios por enfermedad. Todos estos cambios impiden comparar las estadísticas de distintos años, a pesar de que se usan las mismas reglas.

Para solucionar estos problemas de comparabilidad, se introdujo una variable derivada en la versión de 1993 del Registro de Empleo. En el caso de las personas que participaron en la Encuesta sobre Fuerza Laboral (EFL) en noviembre de 1993, los datos de los estados de resultados se combinaron con la situación de empleo de acuerdo con esa misma encuesta. A partir de un análisis de regresión, se construyeron modelos aplicables a diferentes combinaciones de sexo y edad para estos datos de prueba. De esta forma, se asignaron distintos límites de ingresos a los diferentes grupos, pero todos los límites correspondían a las definiciones de empleo de dicha encuesta. Otra ventaja es que las variables administrativas de alta calidad pueden afectar mucho las clasificaciones, a diferencia del reducido efecto de las variables de baja calidad. El siguiente es un esquema general del análisis:

1. A partir de datos de prueba con una categoría de empleo conocida de acuerdo con la EFL, se estiman los modelos de regresión, donde dicha categoría es la variable *y* con dos categorías (empleo remunerado y no empleo remunerado) y los datos sobre estados de resultados son las variables *x* del modelo de regresión. Los datos de prueba se dividen en subgrupos utilizando las categorías de edad y sexo, contenidos en los estados de resultados. Se realizan análisis independientes para cada subgrupo.

2. Con base en el modelo estimado para un subgrupo, se estima un valor *y* usando las variables *x* de los datos incluidos en los estados de resultados. Si el análisis tiene éxito, las personas clasificadas como empleados remunerados en la EFL tendrán valores *y* estimados muy distintos de los valores de aquellas que no son empleados remunerados.
3. Se determina un valor de corte de modo que aquellas personas con valores *y* estimados en un lado del valor de corte se clasifiquen como empleados remunerados y el resto de las personas como no empleados remunerados. Este límite se fija de modo que el número de personas clasificadas como empleados remunerados sea de la misma magnitud que el número correspondiente de acuerdo con la EFL en los datos de prueba.
4. Estos valores de corte para los diferentes subgrupos se usan entonces para que todas las personas incluidas en la población de registro se clasifiquen usando las variables administrativas de los datos consignados en los estados de resultados.

Por ejemplo, para la versión del registro correspondiente a 2001, se usó la EFL de noviembre de 2001 a fin de establecer nuevos límites para los ingresos mediante nuevos análisis de regresión. Estos nuevos límites tienen la misma definición que antes para empleado remunerado en la EFL. De este modo, es posible hacer comparaciones relevantes entre distintos años.

Las variables derivadas cualitativas que se han formado a partir de un modelo estadístico se deben estimar de modo que puedan evaluarse los *errores de clasificación*. Con un buen modelo estadístico, deben minimizarse tanto los *errores netos* como los *errores brutos*. En el gráfico 6.3 se compara el método anterior con el nuevo para definir personas empleadas en el Registro de Empleo.

El *error bruto* estimado en la gráfica 6.3 es una estimación de la proporción de clasificaciones incorrectas en todo el registro, donde la estimación del *error neto* es una estimación del error sistemático en el método de definición de personas asalariadas para el Registro de Empleo (suponiendo que la EFL presenta estimaciones correctas). Para una buena estimación de los errores bruto y neto, es deseable contar con dos conjuntos de datos de prueba, uno para elaborar el modelo estadístico y el otro para estimar los errores de clasificación.

Gráfico 6.3 Errores de clasificación en el Registro de Empleo, 1993

Número de personas en los datos de prueba	Estimación en el **nuevo** Registro de Empleo			Estimación en el **anterior** Registro de Empleo		
	Empleados	No empleados	Total	Empleados	No empleados	Total
Empleados, EFL	22 360	1 158	23 518	22 472	1 046	23 518
No empleados, EFL	1 068	6 872	7 940	1 329	6 611	7 940
Total	23 428	8 030	31 458	23 801	7 657	31 458

Porcentaje del número total de personas	Estimación en el **nuevo** Registro de Empleo			Estimación en el **anterior** Registro de Empleo		
	Empleados	No empleados	Total	Empleados	No empleados	Total
Empleados, EFL	71.1	3.7	74.8	71.4	3.3	74.8
No empleados, EFL	3.4	21.8	25.2	4.2	21.0	25.2
Total	74.5	25.5	100.0	75.7	24.3	100.0

Error de clasificación				
	Error neto:	74.5 – 74.8 = –0.3%	Error neto:	75.7 – 74.8 = 0.9%
	Error bruto:	3.7 + 3.4 = 7.1%	Error bruto:	3.3 + 4.2 = 7.5%

6.2.4 Variables derivadas y valores imputados de variables

Como ya mencionamos, derivar valores de variables por medio de cálculos se relaciona con *imputar* valores de variables. La diferencia es que una variable derivada se usa creando cálculos para *todos* los objetos de un registro, mientras que los valores imputados de variables sólo se calculan para aquellos objetos del registro que no tienen un valor. En el siguiente gráfico, se comparan estos dos tipos de procesamiento.

El gráfico 6.4 muestra un registro longitudinal de empresas con las sumas de los salarios para los años 1 y 2 (*Ssal1* y *Ssal2*), así como el número de empleados en los años 1 y 2 (*Emp1* y *Emp2*). Faltan algunos valores para el número de empleados. Los valores imputados se calculan dividiendo la suma salarial entre el promedio de la suma salarial por empleado.[1] A continuación, se calculan dos mediciones de crecimiento como variables derivadas. *Difss* = 1 si las sumas salariales se han incrementado y *Dife* = 1 si el número de empleados se ha incrementado entre los años 1 y 2.

Gráfico 6.4 Valores imputados y variables derivadas en un registro empresarial

Antes del procesamiento

ID	Ssal1	Ssal2	Emp1	Emp2
1	12 132	12 344	34	32
2	1 775	1 438	5	4
3	893	914	2	perdido
4	18 923	17 835	53	47
5	239	346	perdido	perdido
6	6 221	7 583	17	20
7	549	514	3	2

Con valores imputados y variables derivadas

ID	Ssal1	Ssal2	Emp1	Emp2	Difss	Dife
1	12 132	12 344	34	32	***1***	***0***
2	1 775	1 438	5	4	***0***	***0***
3	893	914	2	***2***	***1***	***0***
4	18 923	17 835	53	47	***0***	***0***
5	239	346	***1***	***1***	***1***	***0***
6	6 221	7 583	17	20	***1***	***1***
7	549	514	3	2	***0***	***0***

6.2.5 Creación de variables mediante codificación

En algunos casos, se usan datos en forma de texto para crear variables estadísticamente útiles. La información se transforma de un texto no estructurado a valores de variables totalmente estructurados en un proceso de codificación.

La dirección donde se ubica un establecimiento es el vínculo entre los registros Inmobiliario y Empresarial. Hay varios problemas con estos domicilios, que actualmente no se presentan en un formato estandarizado. Por ejemplo, la calle del domicilio puede indicarse de diferentes maneras:

Storgatan 17
Storg 17
Storg**.** 17
y también puede haber errores ortográficos: Storgtan 17

Al procesar los domicilios en un programa de traducción, donde las domicilios reales se comparan con listas de alias, muchos de ellos se pueden transformar en un formato estructurado (es decir, todas las variaciones se convierten en Storgatan 17). Los domicilios que no pueden aclararse mediante dicho proceso automatizado pueden codificarse de manera manual. Cuando los domicilios de las propiedades consignados en el Registro Inmobiliario

[1] El valor imputado a la empresa 5 para el número de empleados en el año 1 se calcula a partir de las seis empresas cuyos valores se conocen: suma salarial = 40 493; número de empleados = 114; valores imputados redondeados = 239/(40 493/114) = 1.

y los domicilios de ubicación incluidos en el Registro Empresarial tienen el mismo formato estructurado, los registros se pueden emparejar por medio de la variable de domicilio.

Hay varios ejemplos de variables importantes en el sistema de registros que se crean por codificación. Para algunas empresas, la clasificación industrial se determina usando el directorio telefónico (o páginas amarillas). Las causas de defunción y los tipos de lesiones ocupacionales se codifican usando información en forma de texto de los formularios administrativos.

Causas de defunción

Tras investigar la causa de defunción, un médico elabora el certificado correspondiente. Estos certificados se digitalizan y registran en el sistema de cómputo. Los diagnósticos en estos formularios registrados electrónicamente se codifican usando un programa especial diseñado en la Oficina de Estadística de Suecia. El programa codifica automáticamente los diagnósticos convirtiendo el texto simple en un código acorde con las clasificaciones estadísticas internacionales para causas de defunción. Los casos inciertos se separan y codifican manualmente tras consultar con el médico de turno.

Accidentes de trabajo

Cuando se notifican accidentes de trabajo, cada empleador es responsable de enviar un formulario especial debidamente llenado a la oficina de seguridad social. La autoridad de Medio Ambiente Laboral de Suecia, en este caso la autoridad estadística competente, codifica la descripción en texto libre del accidente de trabajo en diferentes variables estadísticas, como *incidente*, *factor externo principal*, *causa sospechada* y *diagnóstico*. Las instrucciones de codificación, la capacitación de los codificadores y las verificaciones de la codificación son fases importantes del proceso para asegurar la calidad del resultado.

6.3 VALIDACIÓN Y CORRECCIÓN DE LAS VARIABLES DEL REGISTRO

En los capítulos 5 y 6, describimos la manera en que los datos administrativos se transforman en registros estadísticos. Todas las etapas de este proceso incluyen la validación de los datos administrativos, pero como queremos que el término *validación* tenga un significado más preciso, decidimos usar el término validación sólo para el trabajo de validación descrito en esta sección.

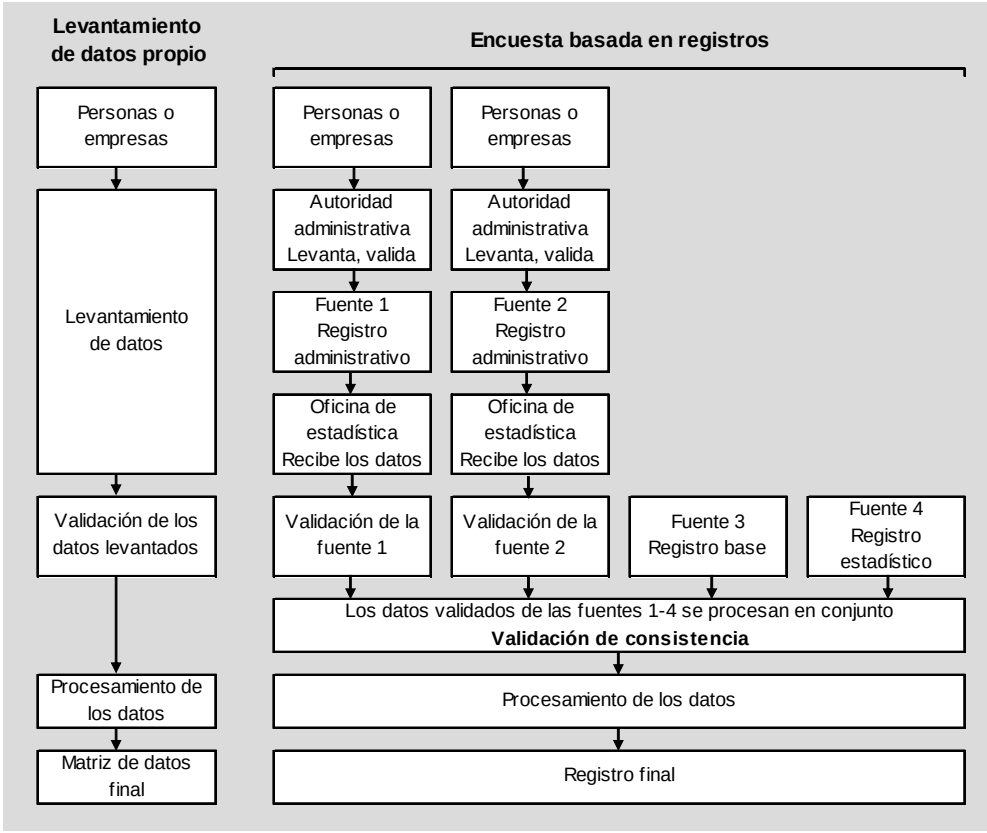
Encuestas con captación de datos propia y encuestas basadas en registros

La principal fase de validación de las encuestas con una captación de datos propia tiene que ver con la validación de los datos recabados. A menudo es posible ponerse en contacto con los proveedores de los datos para corregir los valores de variables que resultan ilógicos. Si la validación requiere una gran cantidad de recursos, puede ser una señal de que es necesario rediseñar el cuestionario. Cabe señalar que los errores o los errores sospechados se interpretan como errores en relación con los valores de las variables. El objetivo consiste en remplazar los valores equivocados o ilógicos por valores imputados corregidos o razonables. En general, en la bibliografía científica sólo se aborda la validación de los datos de encuestas con una captación de datos propia.

En el caso de las encuestas basadas en registros, los datos fueron validados en primera instancia por la autoridad administrativa. Después, se valida cada fuente administrativa cuando los datos se entregan a la oficina de estadística. Pero también hay otra fase más depurada en el proceso de validación, en la que los datos de muchas fuentes (encuestas basadas en registros o censos) se validan conjuntamente. Gracias a esta *validación de la*

consistencia se pueden identificar otros errores e inconsistencias. Se trata de una fase importante que no existe en las encuestas por muestreo con una captación de datos propia. Estos aspectos se ilustran en el siguiente gráfico.

Gráfico 6.5 Validación en encuestas con captación de datos propia y encuestas basadas en registros



Errores en las variables y errores en los objetos

En la validación tradicional de los datos de una encuesta cuya captación ha sido responsabilidad de la oficina de estadística, todos los errores y posibles errores se interpretan como errores en las variables. En la validación de la consistencia de los datos de una encuesta basada en registros, estamos validando datos de distintas fuentes y los errores sospechados pueden ser ocasionados tanto por errores en las variables como por errores en los objetos.

Un error en un objeto significa que creemos estar comparando datos de diferentes fuentes sobre el *mismo* objeto; sin embargo, en realidad los datos que comparamos se refieren a objetos *diferentes*, que erróneamente tienen la misma identidad. Así sucede, por ejemplo, cuando obtenemos coincidencias falsas luego del emparejamiento o cuando no hemos creado correctamente los objetos derivados. En los casos de errores en los objetos, no debemos corregir o imputar los valores de las variables hasta haber verificado que los objetos sean los mismos.

Ejemplo: Definiciones incorrectas de objetos en el Registro Agropecuario (véase también sección 5.4.7)
El Sistema Integrado de Gestión y Control (SIGC) contiene solicitudes de empresas agropecuarias por tipo de subsidio. Wallgren (1999) emparejó este registro con el Registro Agropecuario de la Oficina de Estadística de Suecia según las variables de vinculación NIP y número de organización. Todas las coincidencias se verificaron comparando la misma variable: *superficie de tierras de labranza* en ambos registros. En alrededor de 9% de los casos se detectaron errores pues difería la superficie de tierras de labranza. Estos errores fueron producto de errores en los objetos. Al combinar los objetos con el mismo número telefónico, pudieron derivarse nuevos objetos con la misma superficie de tierras de labranza en las dos fuentes.

Nuestra conclusión es que debemos distinguir entre tipos de validación:

1. Validación común de una fuente donde los errores se interpretan como errores en las variables.
2. Validación de la consistencia de muchas fuentes donde los errores se interpretan como errores en las variables.
3. Validación de la consistencia de muchas fuentes donde los errores se interpretan como errores en los objetos.

El primer tipo de validación se usa para todo tipo de encuestas, mientras que el segundo y el tercero se deben aplicar a las encuestas basadas en registros.

Validación en las encuestas basadas en registros
Una encuesta por muestreo tiene un uso principal y se genera un número limitado de tablas porque el error de muestreo no permitirá una tabulación más desagregada. La validación de los datos se puede reducir para evitar la "sobrevalidación"; se pueden pasar por alto errores que no afectan significativamente las estimaciones finales. Granquist y Kovar (1997) abordan la sobrevalidación. Por su parte, Biemer y Lyberg (2003) ofrecen un panorama general de la encuestas con captación de datos propia.

En muchas encuestas basadas en registros se usa un registro estadístico y se produce un gran número de tablas detalladas. Por consiguiente, resulta difícil definir cuáles son los errores menores que se pueden pasar por alto. Es necesaria entonces la microvalidación, pero se deben diseñar métodos de validación para que el tiempo dedicado a este trabajo sea razonable y mejore la calidad de los datos.

El objetivo de la validación es aumentar la calidad tanto a largo como a corto plazo:

– Detectar y corregir errores. En los datos administrativos puede haber errores obvios que se pueden corregir automáticamente.
– Identificar fuentes de errores y, en cooperación con la autoridad administrativa, reducir su alcance.
– Contribuir a que el personal que trabaja con el registro adquiera más conocimientos sobre el tema de que se trate. El trabajo de validación ayuda al personal a familiarizarse con las características de los datos administrativos, así como a entender cómo se pueden aprovechar los datos.

Para evitar la duplicación del trabajo dentro del sistema de registros, se deben validar las *variables primarias locales* en el registro donde se creó por primera vez la variable dentro del sistema. Después, cuando estas variables se exportan a otros registros, no deberá ser necesaria una nueva validación de cada fuente y se podrá validar la consistencia de manera directa.

Los valores perdidos, o que deben rechazarse, ocasionan una no respuesta parcial. Estos valores perdidos se pueden sustituir por valores imputados. En el capítulo 8 veremos los ajustes por no respuesta en los registros.

A continuación presentamos tres estudios de caso que ilustran el trabajo de validación de diferentes registros en la Oficina de Estadística de Suecia, seguidos por un resumen de los resultados de cada estudio.

6.3.1 Trabajo de validación dentro del Registro de Ingresos y Tributación

Como primer ejemplo de cómo organizar el trabajo de validación, usaremos el Registro de Ingresos y Tributación, que nos servirá para ilustrar la distribución del ingreso y la tributación para individuos y familias con base en las declaraciones de impuestos y otros registros administrativos.

Las variables de este registro también se usan en el modelo de microsimulación FASIT[*]. El Ministerio de Hacienda usa este modelo, entre otros, para estudiar los efectos de los cambios planificados en la tributación y el pago de transferencias. Este registro debe cumplir con muchos requisitos de calidad importantes en un nivel micro, como los valores de los ingresos y la tributación para individuos y familias, que deben estar completos y ser consistentes. Por consiguiente, a fin de asegurar que no se obtengan resultados de simulación extraños, es necesaria una labor exhaustiva de validación y corrección de los valores de las variables.

En la sección 1.4.1 explicamos la creación de un registro usando como ejemplo el Registro de Ingresos y Tributación para ilustrar la manera en que los registros administrativos se transforman en registros estadísticos. En el gráfico 1.4 se detallan los diferentes registros fuente utilizados.

Fase 1 – Validación individual de los registros administrativos

En total, cada año se reciben alrededor de 30 registros administrativos y lo primero que se debe hacer es validar cada uno de ellos como explicamos a continuación.

Para empezar, se verifican las *descripciones de las entradas* incluidas en los nuevos registros administrativos. Por lo general, estas descripciones son distintas cada año, puede haber cambiado el nombre de las variables y se pueden haber agregado nuevas variables. Existe el riesgo de malinterpretar la descripción de las entradas, lo que generaría errores. Por lo tanto, es importante mantener una estrecha comunicación con el personal de la autoridad administrativa que puede dar las explicaciones necesarias.

Después se estudian los *valores extremos* de las variables cuantitativas, que se comparan con los *valores del año anterior*, tanto en el nivel agregado como en el individual. Algunas variables, como el subsidio por enfermedad, pueden tener un valor tope que puede usarse para estas verificaciones. A continuación se llevan a cabo *verificaciones lógicas* para corroborar que los totales presentados sean realmente la suma de sus partes. En todos los casos en que los valores agregados son un monto no esperado, las entradas individuales tienen valores extremos o los totales no coinciden con sus partes, se contacta a la autoridad administrativa para corregirlos. Los errores pueden ser producto de interpretaciones equivocadas de las descripciones de las entradas o quizá los valores extremos obedecen a una medida tributaria que no debe afectar las estadísticas. En el caso de las entradas sobre individuos, la

[*] Nota de la traductora: FASIT es la abreviación en sueco de Sistema de Análisis de la Distribución de Ingreso y Transferencias.

frecuencia de errores es tan baja que es posible hacer correcciones registro por registro en consulta con la autoridad administrativa.

En ocasiones, algunas autoridades subordinadas comunican ciertas variables a la autoridad central, que a su vez entrega los datos a la Oficina de Estadística de Suecia. En esos casos, conviene verificar que todas las autoridades subordinadas hayan entregado datos. Por ejemplo, cada año normalmente se pierden datos sobre asistencia social para algunos municipios. En el Registro de Ingresos y Tributación se debe documentar qué municipios no han aportado datos. Se pueden hacer imputaciones para estos valores perdidos, por lo general usando los datos de asistencia social del año anterior.

La validación no sólo tiene un efecto directo en la calidad del registro. Si se organiza de tal modo que varias personas compartan la responsabilidad, la validación puede contribuir a la cooperación y el intercambio de experiencias dentro del equipo de trabajo. De este modo, aumentan los *conocimientos sobre la materia de que se trate* e, indirectamente, incluso la calidad del registro. Después, cuando se documenta el trabajo y se toman medidas para corregir los datos, los conocimientos sobre la materia se fortalecen aún más.

El *contacto con los proveedores de datos* tiene varios efectos importantes. En primer lugar, se debe explicar al personal de la autoridad administrativa cómo y con qué propósito se usan sus datos en la Oficina de Estadística de Suecia. Ese personal debe entender las consecuencias que tiene la falta de calidad de los datos para los usuarios de las estadísticas. A fin de fomentar que el personal de la oficina de estadística adquiera un mayor conocimiento sobre la materia, también es importante el contacto con los proveedores de datos. Por eso, el personal que trabaja con el Registro de Ingresos y Tributación sostiene reuniones con el Consejo Nacional Tributario dos veces al año. Este contacto también es de utilidad para identificar nuevas fuentes administrativas.

En principio, no es necesario que el personal del Registro de Ingresos y Tributación verifique de nuevo los registros que usa como fuentes dentro del sistema de registros de la Oficina de Estadística de Suecia, pues las verificaciones debieron hacerse en el registro primario.

Algo que hasta la fecha no ha funcionado particularmente bien es el intercambio de experiencias entre las distintas unidades de la Oficina de Estadística de Suecia. ¿Cuántos funcionarios saben exactamente qué datos se captan para el Registro de Ingresos y Tributación? ¿Aquellos funcionarios que tienen datos que podrían servir para las necesidades de ese registro saben cuáles son sus requerimientos? Por ejemplo, ¿el Registro Tributario para Bienes Inmuebles cumple con los requerimientos que imponen los cálculos de impuestos y patrimonio para el Registro de Ingresos y Tributación? La falta de cooperación no debe hacer que varias unidades dentro de la Oficina de Estadística de Suecia reciban y verifiquen los mismos datos administrativos. Se debe evitar la duplicación del trabajo tanto para la Oficina de Estadística de Suecia como para las autoridades que proporcionan los registros.

Fase 2 – Verificación final de todo el registro

En la fase anterior, se verifican todos los datos de cada autoridad. En esta fase, se combinan todas las variables de todas las fuentes en un solo registro total, de modo que se puedan comparar las distintas fuentes. Entonces se generan todas las variables derivadas. De este modo, se pueden realizar nuevas verificaciones de la consistencia, es decir, corroborar que la suma de todos los valores de las variables de los diferentes registros concuerde con la suma de otra fuente. Así se pueden detectar errores adicionales.

Esta etapa del trabajo de validación se lleva a cabo en *una muestra de las entradas* de todo el registro. Esto obedece en parte a que de otro modo el trabajo sería demasiado engorroso y extenso, y en parte a que hay requisitos de calidad particularmente estrictos para el

segmento del registro total que se usará en el modelo FASIT. El registro total consta de alrededor de nueve millones de entradas con 500 variables, razón por la cual el método de muestreo es un requisito previo para una verificación precisa. Cabe señalar que el proceso de validación permite al personal aumentar sus conocimientos sobre la materia —conocer los datos en el nivel micro contribuye a desarrollar competencias importantes— y trabajar con una muestra puede ser una buena manera de familiarizarse con los datos.

Usualmente, no se identifican tantos errores de esta manera, de modo que se pueden corregir de manera manual consultando a las autoridades administrativas. Si las correcciones en la muestra no afectan tanto las estimaciones, se puede aprobar el registro completo y dar por terminada la fase de verificación. Si las correcciones de la muestra son de mayor envergadura, podrían ser necesarias algunas verificaciones y correcciones automáticas de todos los elementos del registro. Sin embargo, actualmente no se usan estos métodos.

Fase 3 – Verificación de estimaciones

En esta fase, se elaboran todas las tablas importantes usando como base el registro completo. Las estimaciones se verifican y comparan con los valores del año anterior. Además, se realizan varias simulaciones con el modelo FASIT, con la única finalidad de poner a prueba la calidad de los datos. Si, por ejemplo, el subsidio para vivienda no sufre cambios en el modelo, entonces éste debe generar valores que concuerden con las tablas elaboradas anteriormente.

6.3.2 Validación dentro del Registro de Estados de Resultados

El Registro de Estados de Resultados se usa para calcular las sumas salariales por rama industrial específica y también cuando se crean los registros de Actividades y de Empleo. En esta sección explicamos el trabajo de validación realizado con los datos definitivos incluidos en los estados de resultados para el ejercicio fiscal 2001, que la Oficina de Estadística de Suecia recibió en octubre de 2002. Los responsables del Registro de Estados de Resultados verifican los datos de los estados de resultados y el registro validado se usa como fuente para otros registros dentro del sistema.

Verificación de las definiciones de población

El *primer paso* en el proceso de validación consiste en verificar que el número de estados de resultados recibidos concuerde con los enviados por el Consejo Nacional Tributario.

El *segundo paso* es crear una matriz con los datos finales de los estados de resultados de acuerdo con todas las modificaciones en el paquete. El Consejo Nacional Tributario no cambia los datos de entrada, sino que el proveedor de los datos (en este caso el empleador) remite las modificaciones y se crean nuevas entradas que suponen la eliminación, modificación o sustitución de entradas anteriores. Por ello, es necesario procesar el registro para descartar las entradas no válidas y verificar los registros duplicados. Los valores de variables para alrededor de 10 300 entradas se trasladan del estado de resultados original a la entrada con la modificación, pues ésta puede estar incompleta.

El *tercer paso* de la validación consiste en verificar todas las identidades. Como los estados de resultados pueden contener identidades individuales y empresariales, se deben verificar tanto los números de identificación personal como los números de organización. Alrededor de 7 600 números de identificación personal estuvieron incorrectos, de los cuales 5 000 pudieron corregirse automáticamente.

El *cuarto paso* es el emparejamiento de los números de identificación personal (NIP) incluidos en los estados de resultados con los del Registro de Población para el 31/12/2001 y el emparejamiento de las identidades empresariales con el Registro Empresarial para

marzo de 2002. En ambos casos, se identifican varias entradas no emparejadas: el Registro de Estados de Resultados contiene más NIP que el Registro de Población y aproximadamente 12 200 identidades empresariales más que el Registro Empresarial. Como los datos de los estados de resultados se relacionan con el año completo, se deben emparejar con las *versiones por año calendario* de los registros de Población y Empresarial. Las 12 200 entradas que no aparecían en el Registro Empresarial son sobre todo identidades usadas de manera temporal para empresas en quiebra, etc y, hasta cierto punto, de NIP incorrectos.

Verificación de los valores de las variables

En el *quinto paso*, se verifican los errores de desviación usando 16 verificaciones distintas de probabilidades. La relación entre ingreso percibido e impuestos se usa de distintas maneras, y además se hace una búsqueda de las entradas con ingresos percibidos o impuestos extremadamente altos. En estas verificaciones, se identifican alrededor de 5 000 entradas con valores extremos, las cuales se verifican de manera simple y sólo unas cuantas con el Consejo Nacional Tributario. Después de estos pasos, cada estado de resultados se considera aceptado, se sustituye por un nuevo estado o se elimina del registro.

Verificación de la variable más importante

La fase más importante de la validación consiste en verificar que las personas empleadas estén vinculadas con los establecimientos correctos. Este vínculo es crucial para todo el sistema del registro porque permite dar cuenta de las personas empleadas en trabajos remunerados por rama industrial y región. Las dificultades con este vínculo se presentan cuando las empresas tienen más de un establecimiento. A pesar de que el empleador tiene el deber de indicar el establecimiento en cada estado de resultados, muchas veces falta esta información y en ocasiones no es aceptable. Los números de establecimiento no aceptables se identifican comparando el número de empleados con los datos correspondientes en el Registro Empresarial y con los datos de la versión del año anterior del Registro de Estados de Resultados. También se considera si la distancia de desplazamiento entre el lugar de trabajo y el de residencia es aceptable.

Cuando en el estado de resultados de una empresa con más de un establecimiento, falta un establecimiento o hay alguno que parece no aceptable, se contacta al empleador para hacer una captación de datos especial mediante un cuestionario de actualización del registro. Los responsables del Registro de Estados de Resultados trabajan con los responsables del Registro Empresarial para identificar cambios relacionados con el código municipal del establecimiento y el código de clasificación industrial.

Validación de resultados

El Registro de Estados de Resultados se usa como fuente para el Registro de Empleo, de modo que al verificar los resultados de este último también se verifica la calidad del primero. Las tablas desagregadas con personas empleadas por rama industrial y municipio se reúnen y comparan con las tablas del año anterior. Se verifican las desviaciones y se documentan los resultados de estas verificaciones. Esta documentación es muy útil porque muchos usuarios hacen preguntas luego de la publicación de resultados y los cuestionan. Cuando existe documentación, los funcionarios que están en contacto con los usuarios pueden responder: “Hemos verificado y los resultados son correctos hasta donde alcanzamos a ver”.

6.3.3 Validación de las declaraciones de ingresos de las empresas

El Registro de Empresas y Base de Datos de Personas Físicas (FRIDA, por sus siglas en inglés) se apoya en las declaraciones de ingresos de las empresas con variables tributarias.

Este registro se usa de manera similar al modelo FASIT presentado en la sección 6.3.1, pero el objetivo es estudiar los efectos tributarios en las empresas y los trabajadores por cuenta propia. El FRIDA es producto de varios años de estrecha cooperación con el Ministerio de Hacienda. Se trata de un registro documentado —lo que incluye los tipos de cuestionarios administrativos— con una metodología de validación descrita.

La idea original era crear un registro que incluyera a todas las empresas, pero se observó que la calidad de los datos era demasiado baja para ello. El trabajo de validación habría sido demasiado oneroso para crear un registro completo de buena calidad. Con el paso de los años, no sólo ha mejorado la calidad de los datos, también se ha diseñado un programa de validación tan avanzado que ahora sí podría considerarse la creación de un registro completo. Uno de los requisitos indispensables para el uso de este registro en modelos de simulación es que la validación debe ser tan eficaz que se obtengan datos de alta calidad. La gran demanda de calidad es una de las razones por las que se ha decidido utilizar los datos definitivos de las declaraciones de ingresos y compararlos con los datos sobre las decisiones tomadas por el Consejo Tributario, los cuales están disponibles al mismo tiempo. Dada la complejidad de los datos, se ha dedicado largo tiempo a la adquisición de conocimientos sobre la materia y a mantener un contacto cercano tanto con la autoridad administrativa como con el cliente.

La primera fase de este trabajo consiste en recibir el paquete de datos de la autoridad administrativa, en este caso el Consejo Nacional Tributario. El paquete es grande y los datos no tienen la estructura que se usa habitualmente en la Oficina de Estadística de Suecia, sino que incluye diversas subentradas: *identidad, tipo de entrada* y *valor*. Los datos se envían por variable y no por renglón u objeto. La transferencia de los datos, que consisten en alrededor de un millón de cuestionarios con 300 a 400 variables, puede tardar una semana.

En la siguiente fase todos los datos se reestructuran en matrices, donde todas las variables relacionadas con el cuestionario administrativo de una empresa se incluyen en un renglón. Se crea una matriz de datos por cada tipo de formulario o cuestionario administrativo. Son necesarias verificaciones minuciosas para corroborar que todos los datos se incluyan e interpreten correctamente en el momento de su reestructuración. Los formularios y cuestionarios administrativos cambian de un año a otro, por lo que se requiere mantener contacto con el Consejo Nacional Tributario y tener acceso a todos los tipos de cuestionarios para el año. Estos cuestionarios son complejos pues incluyen muchos conceptos económicos y tributarios, de modo que la unidad que recibe las declaraciones de impuestos de las empresas necesita un alto nivel de conocimientos especializados para interpretar y usar correctamente el material.

Después de recibir y estructurar los datos como acabamos de describir, en la fase siguiente se crean *unidades empresariales derivadas* de propietarios individuales. Es difícil trabajar con este grupo de empresas, ya que incluye varios tipos de cuestionarios y muchos trabajadores por cuenta propia con diversas actividades. Todas las empresas que se pueden vincular con personas físicas dentro de un hogar con un mismo perfil tributario conforman una unidad empresarial.

Se hace una primera verificación en todo el conjunto de datos para detectar errores técnicos específicos. Uno de esos errores puede ser que el número de recuadro en el cuestionario se haya pegado a la cantidad. Si, por ejemplo, al lado del recuadro 605 se indica como resultado 1 000 000 coronas, esta cantidad podría volverse incorrectamente 6051000000. Por consiguiente, el programa de verificación debe ser capaz de reconocer la secuencia “605”. Después estos errores se corrigen manualmente.

La población de empresas se estratifica por tipo y tamaño y, a continuación, se selecciona la muestra para el FRIDA. Las empresas inactivas se eliminan y las entradas no definidas como inactivas pasan por el programa de verificación y corrección automáticas.

Los totales incorrectos, los errores de captura y los cambios tributarios que no se han corregido en los datos del Consejo Nacional Tributario son errores comunes. El programa de verificación y corrección revisa todos los totales y también la concordancia de los valores que se han trasladado de un recuadro a otro del cuestionario. Los diferentes tipos de errores detectados durante el año en el FRIDA se han incluido en el programa y la corrección se efectúa automáticamente. Si un total concuerda luego de una modificación para un tipo de error específico, entonces la corrección se hace de acuerdo con ese tipo de error. De esta manera se corrigen los errores en los símbolos, en el número de ceros, etc.

Una submuestra consiste en alrededor de 40 000 empresas de responsabilidad limitada. Aproximadamente 30 000 de ellas pasan por el programa de verificación sin cambios, pues los totales concuerdan. Pero se detectan errores en alrededor de 10 000 empresas, 8 000 de las cuales se pueden corregir automáticamente usando todos los tipos de errores que contiene el programa. Las 2 000 restantes deben pasar por una nueva verificación. Alrededor de 1 700 tienen errores menores, desde el punto de vista del porcentaje, de modo que se aprueban sin mayores modificaciones. Por lo tanto, quedan en torno de 300 empresas con errores considerables que se deben verificar manualmente. Estas empresas se comparan con su respectivo informe anual, presentado a la Oficina de Patentes y Registros. En ese momento se pueden corregir las entradas para alrededor de 60 empresas, pero quedan 240 que no. Por cada total incorrecto, se crea una variable específica donde se almacena la diferencia. A cada una de las correcciones se le asigna un código de corrección. A las entradas con errores que no se pueden corregir se les asigna un código de error.

La reglamentación tributaria va cambiando, de modo que los datos cambian cada año, desaparecen algunas variables y se agregan otras nuevas. Esto implica que el programa de validación se debe modificar regularmente. También se modifican las necesidades de los usuarios, lo que significa que cualquier cambio en la validación se debe analizar junto con ellos para que la Oficina de Estadística de Suecia se concentre en los problemas más importantes. Como el personal de esa dependencia "convive con datos en el nivel micro", el proceso de aprendizaje será continuo, lo que favorecerá un mayor dominio de la materia. Es importante que el personal también use los datos y analice regularmente los problemas de los usuarios.

6.3.4 Resumen: procesos de validación en el sistema de registros

Al analizar los tres casos, observamos que la validación de los datos de las fuentes administrativas se puede estructurar de acuerdo con las siguientes seis fases. En los recuadros describimos brevemente los casos antes mencionados:

1. Se crea una matriz de datos y *se combinan todas las entradas* correspondientes al mismo objeto.

Registro de Estados de Resultados:
Se crea un matriz de datos con los datos finales de los estados de resultados después de recibir todos los cambios en el paquete. Por ello, es necesario procesar el registro para eliminar las entradas no válidas y verificar que no haya entradas duplicadas.

2. Se debe llevar a cabo una validación para *verificar el registro de población*, es decir, que no haya objetos perdidos, ni externos o duplicados. Esto se hace mediante un cotejo con el

registro base relevante y también incluye verificar todas las variables de identidad, incluido su formato.

Registro de Estados de Resultados:
Alrededor de 7 600 números de identificación personal (NIP) fueron incorrectos, 5 000 de los cuales pudieron corregirse automáticamente. Los NIP asentados en los estados de resultados se emparejaron con el Registro de Población correspondiente al 31/12/2001 y la identidad de las organizaciones con el Registro Empresarial para marzo de 2002. En ambos casos hubo muchas entradas no emparejadas.

3. La validación puede incluso ser necesaria para *verificar el objeto*, es decir, corroborar que los datos correspondientes a una identidad específica de diferentes fuentes se refieran al mismo objeto. Esto se puede hacer comparando valores para variables similares de fuentes diversas.

FRIDA, declaraciones de ingresos de las empresas:
Se crean unidades empresariales derivadas para los individuos propietarios de empresas. Todas las empresas que se pueden vincular a personas dentro de un hogar con el mismo perfil tributario constituyen una unidad empresarial.

Sección 6.3, primer ejemplo:
El registro SIGC se emparejó con el Registro Agropecuario de la Oficina de Estadística de Suecia. Todas las coincidencias se verificaron comparando la misma variable, a saber, la superficie de tierras de labranza en ambos registros. En aproximadamente 9% de los casos, se detectaron errores causados por errores en los objetos. Pudieron derivarse nuevos objetos para los que la superficie de tierras de labranza era la misma en ambas fuentes.

4. Debe verificarse que la *entrega sea completa* en lo que respecta a los objetos y las variables. Muchas variables administrativas sólo se proporcionan para los objetos de que se trate; si el valor es cero, no hay datos administrativos. Por ello resulta difícil diferenciar entre valores perdidos y verdaderos valores cero. Una entrega incompleta de variables se investiga verificando las tablas en las que los totales o frecuencias se comparan con los valores del año anterior. Los totales o frecuencias se calculan para grupos adecuados, como municipios, de modo que puedan detectarse las entregas incompletas.

Registro de Estados de Resultados:
El primer paso del proceso de validación consiste en verificar que el número de estados de resultados recibidos concuerde con los enviados al Consejo Nacional Tributario.

Registro de Ingresos y Tributación:
Por ejemplo, cada año suele haber datos perdidos sobre asistencia social. Se debe documentar en el Registro de Ingresos y Tributación qué municipios no han proporcionado datos a fin de hacer imputaciones para esos valores perdidos. Por lo general, para ello se usan los datos sobre asistencia social del año anterior.

5. Se deben llevar a cabo verificaciones para *revisar los valores de las variables*, de modo que los errores obvios o sospechados puedan detectarse y corregirse. Las descripciones de registros mal interpretadas o los errores técnicos pueden dar origen a *errores obvios* fáciles de corregir.

Registro de Ingresos y Tributación:
Las descripciones de entradas suelen cambiar cada año, pues pueden haber cambiado los nombres de las variables y haberse agregado nuevas variables. También puede mal interpretarse la descripción de los registros, lo que generará errores.
Se estudian los valores extremos de las variables cuantitativas, comparándolos con los valores del año anterior.
Se efectúan verificaciones lógicas para corroborar que los totales realmente correspondan a la suma de sus partes. Los errores pueden ser producto de una interpretación equivocada de la descripción de las entradas y los valores extremos pueden obedecer a una medida tributaria que no debería

afectar las estadísticas.
Verificación de las estimaciones preparadas: en esta etapa, se forman todas las tablas importantes usando el registro completo como base. Las estimaciones se verifican y comparan con los valores del año anterior.

Registro de Estados de Resultados:
Los errores de desviación se verifican mediante 16 verificaciones distintas de probabilidades. La relación entre el ingreso percibido y los impuestos se usa de varias maneras, además de que se buscan las entradas con un ingreso percibido o impuestos extremadamente elevados. Gracias a estas verificaciones, se detectaron alrededor de 5,000 entradas con valores extremos.

Luego de estas etapas de verificación, cada estado de resultados se da por aceptado, se sustituye por un nuevo estado de resultados o se elimina del registro.

Se integran tablas detalladas con empleados por rama industrial y municipio, y se comparan con las tablas del año anterior. Se verifican las desviaciones y se documentan los resultados de estas verificaciones.

6. Se debe *documentar* el proceso de validación, informando en primer lugar la metodología aplicada y después el alcance de los diferentes tipos de error y cómo se corrigieron. Los resultados de las verificaciones se deben documentar especificando los códigos tanto de error como de corrección del registro completado e incluyendo una descripción del efecto y el alcance de las verificaciones.

Registro de Estados de Resultados:
Se integran tablas detalladas con empleados por rama industrial y municipio, y se comparan con las tablas del año anterior. Se verifican las desviaciones y se documentan los resultados de estas verificaciones.

Esta documentación es muy útil porque muchos usuarios hacen preguntas luego de la publicación de los resultados y los cuestionan. Cuando existe documentación, los funcionarios que están en contacto con los usuarios pueden responder: “Hemos verificado y los resultados son correctos hasta donde alcanzamos a ver”.

FRIDA, declaraciones de ingresos de las empresas:
Por cada total incorrecto, se crea una variable específica donde se almacena la diferencia. A cada una de las correcciones se le asigna un código de corrección. A las entradas con errores que no se pueden corregir se les asigna un código de error. Se trata de un registro documentado —lo que incluye los tipos de cuestionarios administrativos— y con una metodología de validación descrita.

6.3.5 ¿Qué más podemos aprender de estos ejemplos?

Los ejemplos anteriores muestran que los datos administrativos recibidos en la Oficina de Estadística de Suecia pueden contener errores que requieren una verificación en el nivel micro. Una vez detectados, estos errores normalmente se corrigen de manera simple. Las necesidades que debe satisfacer el proceso de verificación dependerán del uso que se dará al registro. Los registros estadísticos de esta dependencia suelen utilizarse para fines de investigación; para ese tipo de necesidades analíticas avanzadas, la calidad en el nivel micro deberá ser mayor que cuando sólo se producen tabulados simples y las verificaciones serán más estrictas. Estas exigencias se aplican sobre todo a los estudios longitudinales, donde los vínculos en el nivel micro deben ser razonables.

Conocimiento de la materia y contactos con proveedores
Una conclusión general es que el conocimiento de la materia reviste gran importancia para la eficacia de la validación y las verificaciones. En el caso de encuestas con una captación de datos propia, basta con familiarizarse con la encuesta de que se trate, que normalmente no cambia. No obstante, cuando hablamos de encuestas basadas en registros, será necesario familiarizarse con el sistema administrativo que generó los datos. Dicho sistema puede contener muchas variables complicadas que cambian con frecuencia.

La validación no sólo tiene un efecto directo en la calidad del registro. Si se organiza de tal modo que varias personas compartan la responsabilidad, puede fomentarse la cooperación y el intercambio de experiencias dentro del equipo de trabajo. De este modo, aumentan los *conocimientos sobre la materia de que se trate* e, indirectamente, incluso la calidad del registro. Después, cuando se documenta el trabajo y las medidas tomadas para corregir los datos, los conocimientos sobre la materia se fortalecen aún más.

El método de trabajar a fondo con una muestra de un registro grande podría aplicarse de manera más generalizada. En el caso del Registro de Ingresos y Tributación, es natural trabajar con la muestra usada en el modelo de microsimulación FASIT, pero otros registros también podrían usar este método para poner a prueba la calidad y conocer más a detalle los datos y los temas con los que se relacionan, dicho análisis no sería posible con el registro completo.

Este ejemplo también muestra la importancia tanto de la cooperación y la adquisición de conocimientos dentro del equipo de trabajo que recibe los registros administrativos como de una buena comunicación con las autoridades que proporcionan los datos. Además, se debe estimular la cooperación entre los diferentes equipos que trabajan con registros relacionados para un aprovechamiento eficaz de los datos administrativos.

Si el personal de la oficina de estadística "convive con los datos en el nivel micro", el proceso de aprendizaje será continuo, lo que propiciará un mayor dominio de la materia. Este proceso de aprendizaje se fortalece mediante el contacto estrecho con los usuarios.

Duplicación del trabajo e intercambio de experiencias

Puede duplicarse el trabajo de validación; varias unidades de la Oficina de Estadística de Suecia podrían validar conjuntamente las declaraciones de ingresos. También puede haber un trabajo conjunto en relación con los datos de evaluación inmobiliaria. Como la validación de los datos de registro aún no ha atraído gran atención, no se ha producido este intercambio de experiencias, que resulta esencial a fin de mejorar este aspecto de la metodología para el aprovechamiento de los registros estadísticos.

Programas automatizados de validación e imputación

Se deben diseñar rutinas de validación automatizadas a fin de detectar y corregir errores en nuestros vastos registros. Es necesario formular métodos para validar y corregir microdatos en registros grandes, pues los métodos para la macrovalidación de encuestas por muestreo no satisfacen todas las necesidades de las estadísticas basadas en registros. El caso del FRIDA (declaraciones de ingresos de las empresas) puede servir como ejemplo. El programa de validación y corrección que se usa en ese registro se basa en un amplio conocimiento de la materia adquirido continuamente.

Dentro del proyecto EUREDIT de la Unión Europea, descrito por Pannekoek y de Waal (2005), se han diseñado métodos de validación e imputación completamente automatizados. Falta por ver si este tipo de programas de cómputo se pueden usar para validar datos administrativos. Sin embargo, la automatización no debe obstaculizar nunca la adquisición de conocimientos en la materia de que se trate.

De acuerdo con de Waal y Quere (2003), sólo unos cuantos países (entre los que no figura Suecia) han elaborado rutinas de validación avanzadas que validan y hacen imputaciones automáticamente con base en el paradigma de Fellegi-Holt (1976). Estos métodos se pueden usar para validar de manera individual las fuentes administrativas, pero la validación simultánea de la consistencia de muchas fuentes, mencionada al inicio de la sección 6.3, genera más problemas metodológicos porque pueden generarse errores causados por errores en las variables y los objetos.

Ejemplo: Validación e imputación automáticas de las declaraciones de ingresos de las empresas.

Este ejemplo ilustra la validación e imputación automáticas, que se usan cuando se hace la primera validación de los datos administrativos con una fuente a la vez. En total, se recibieron 464 567 declaraciones de impuestos de pequeñas empresas correspondientes a 2004, que deben presentar un formulario fiscal especial.

Se diseñó un programa de cómputo especial para este formulario fiscal, cuyos métodos se basan en el conocimiento de la materia. Las últimas dos variables de la tabla siguiente, "Ajustes" e "Ingresos gravables", tienen la calidad más alta pues son de gran importancia en materia legal y fiscal administrativa. El programa comienza por estos valores y después busca errores en los signos positivos o negativos y las sumas. De esta forma, es bastante fácil identificar y corregir la mayoría de los errores en este registro de datos de empresas.

En el gráfico 6.6 vemos que los datos son inconsistentes antes de la validación:

Percepciones – Costos – Depreciaciones, etc.
no suman los ingresos totales
(40 734 –665 016 – 2 877 ≠ – 8 825)

La principal causa de estos errores es el costo C5; si se corrigen los errores de siete entradas en esta variable, desaparecerán casi todas las inconsistencias en este registro de 464 567 entradas.

También se deben corregir los costos C3 y C4; 16 correcciones harán que en los costos haya una diferencia de casi 13 mil millones.

Hay 12 855 errores menores en la variable D6. En este caso no importa si se corrigen o no, pues la suma total de D6 casi no cambia.

Gráfico 6.6 Validación e imputación automáticas

		Antes de la validación Miles de millones (coronas)	Después de la validación Miles de millones (coronas)	Número de correcciones
Percepciones				
	P1	31.017	30.793	115
	P2	9.323	9.315	23
	P3	0.394	0.392	4
Percepciones totales		**40.734**	**40.500**	
Costos				
	C1	-8.845	-8.833	22
	C2	-0.913	-0.913	0
	C3	-10.363	-0.961	3
	C4	-6.871	-3.110	13
	C5	-628.046	-3.006	7
	C6	-2.254	-2.252	6
	C7	-7.725	-7.602	23
Costos totales		**-665.016**	**-26.676**	
Depreciaciones, etc.				
	D1	-4.078	-4.075	25
	D2	3.880	3.339	4
	D3	-3.217	-3.216	3
	D4	-0.975	-0.957	1
	D5	0.905	0.905	1
	D6	0.607	0.751	12 855
Depreciacones, etc. totale		**-2.877**	**-3.252**	
Ingresos		**-8.825**	**10.572**	
Ajustes		**-17.789**	**-17.628**	
Ingresos gravables		**-7.053**	**-7.056**	

Las conclusiones de este ejemplo son que:

- Los datos administrativos se deben validar y es recomendable usar un programa informático automático cuando hay muchas entradas.
- En muchos casos, un pequeño número de errores graves puede destruir los datos. En términos generales, es fácil detectar y corregir estos errores.
- Se deben aplicar los conocimientos en la materia al diseñar el programa de cómputo.

Registros base: no existe una versión adecuada

Los tres casos que acabamos de describir requieren *versiones de año calendario* del Registro de Población o el Registro Empresarial para definir sus poblaciones de registro. Estas versiones de los registros no existían cuando se hizo la validación en estos casos, lo que causó no emparejamientos innecesarios. De acuerdo con los principios enunciados en la sección 5.4, se debe crear la mejor versión posible de año calendario usando todas las fuentes disponibles una vez que se han entregado las declaraciones de ingresos de las

empresas. La versión de año calendario debe contener todas las empresas que han tenido actividades durante el año de referencia.

Posible necesidad de captar datos adicionales

Cuando se observa que una variable de los datos administrativos tiene un bajo nivel de calidad para propósitos estadísticos, puede ser necesario captar datos adicionales a fin de lograr un nivel de calidad suficiente. Un ejemplo de ello es la validación que se lleva a cabo cuando se verifica que las personas empleadas estén vinculadas con los establecimientos correctos en el Registro de Estados de Resultados. Para conseguir un nivel de calidad suficiente, se entra en contacto con algunos empleadores mediante una captación especial de datos con un cuestionario de actualización del registro.

La validación como base para la declaración de calidad

Si el personal del registro "convive con datos en el nivel micro", tendrá buena idea de la calidad del registro. Este conocimiento representa una base sólida para redactar las declaraciones de calidad.

6.4 CREACIÓN DE REGISTROS LONGITUDINALES

Kardaun y Loeve (2005) describen y comparan los análisis longitudinales realizados en algunas oficinas de estadística. Comentan que las encuestas longitudinales basadas en registros son comunes en los países escandinavos, mientras que las encuestas basadas en la captación de datos son comunes en los países anglosajones. También señalan que la mayoría de las encuestas longitudinales están orientadas a las personas y sólo unas cuantas a las empresas.

En la sección 1.6, mencionamos que las encuestas basadas en registros tienen la ventaja de una cobertura completa del tiempo, por lo que son adecuadas para el análisis longitudinal. No obstante, para aprovechar de esta manera los datos administrativos antes se debe crear el registro longitudinal, de modo que la calidad longitudinal de los objetos y variables sea suficientemente buena.

En el gráfico 2.10, se ilustra el sistema de registros estadísticos. Aparecen cinco registros longitudinales sobre personas y uno sobre empresas o establecimientos.

¿Para qué se usa el registro estadístico? El uso del registro determina los requisitos de calidad que se deben cumplir y ése es el factor determinante para el tipo de validación que se necesita y el tipo de procesamiento de registro que se debe aplicar.

Un registro que sólo se usa para producir estadísticas oficiales anuales debe proporcionar estimaciones de buena calidad en un nivel agregado. Sin embargo, si el registro también se usará para análisis longitudinales, es necesario definir los *objetos* o *unidades estadísticas* para que se les pueda dar seguimiento a lo largo del tiempo sin que ocurran cambios, sin un significado estadístico, que alteren las trayectorias.

Ejemplo: Si una empresa modifica su situación jurídica y deja de ser una sociedad por acciones para convertirse en una empresa de responsabilidad limitada, no deberá interpretarse como el cierre de una empresa y la constitución de una nueva. Un nuevo producto estadístico de la Oficina Estadística de Suecia denominado Dinámica de Empresas y Establecimientos, tiene la finalidad de presentar un mapa más detallado de los cambios estructurales en el sector empresarial. El objetivo de ese registro es mejorar la clasificación de los eventos dentro de las empresas, para así ofrecer nuevas posibilidades que permitan estudiar la movilidad de la fuerza laboral.

Sobre todo, se crean identidades estables para las empresas, lo que nos permite darles seguimiento con el paso del tiempo de una manera más analítica de lo que permite el Registro Empresarial. Con el registro Dinámica de Empresas y Establecimientos, se da seguimiento simultáneo a las empresas recién constituidas, los cierres, las divisiones y las fusiones. Se considera que el principal indicador de cambio es el personal de la empresa en diferentes momentos, mientras que se otorga una menor importancia al cambio de propietario, rama industrial o ubicación.

La regla básica en este registro es que si la mayoría de los empleados en el año 1 constituyen también la mayoría de los empleados en el año 2, las unidades para ambos años se consideran como la misma empresa, independientemente del número de la organización. El resultado es que al establecer otras reglas para el flujo de personal entre los dos años, es posible clasificar las fusiones y divisiones de empresas.

Del mismo modo, también se deben definir las *variables* para que los cambios con el paso del tiempo tengan un significado estadístico. Por ejemplo, cuando las reglas administrativas cambian en el área tributaria, las variables afectadas no serán comparables a lo largo del tiempo.

Ejemplo: Las reglas para gravar las utilidades declaradas por las empresas de responsabilidad limitada han cambiado varias veces. Estas reglas son importantes porque determinan si el propietario de la empresa debe considerar un superávit como utilidades o como salario, algo que afecta a las estadísticas que incluyen sumas salariales y superávits operativos. Por consiguiente, las variables administrativas *salario para el propietario de la empresa* y *utilidades declaradas* se deben combinar en una variable que tenga un significado estadístico para estas empresas de responsabilidad limitada.

Incluso las variables formadas dentro de una oficina de estadística pueden ser más o menos apropiadas para los análisis longitudinales. El método para formar la variable *personas en empleos remunerados en noviembre* en el Registro de Empleo, según se describe en la sección 6.2.3, puede ocasionar cambios ficticios en las personas ubicadas cerca de los límites de ingresos usados en el modelo estadístico, que constituye la base de la variable derivada.

Gráfico 6.7 Cambios ficticios en la categoría de empleo de una persona

	Año 1	Año 2	Año 3	Año 4	Año 5
Ingreso anual por persona	47 400	48 500	49 600	52 800	53 900
Límite anual en el modelo	47 000	48 600	49 400	52 100	54 200
Categoría de empleo, valor estimado	Empleado	No empleado	Empleado	Empleado	No empleado

Con este tipo de variable derivada en el Registro de Empleo, se pueden producir estimaciones en un nivel agregado para permitir comparaciones de buena calidad con el paso del tiempo. Sin embargo, si se usa la misma variable para los análisis longitudinales, donde se estudia el cambio de categoría de empleo en un grupo de personas, los cambios ficticios pueden significar que la calidad no es suficientemente buena. Puede obtenerse una estimación de buena calidad de la proporción de personas empleadas, pero quizá no sea posible estimar con una calidad suficiente la proporción de personas que han cambiado de categoría de empleo.

En los usos avanzados de los registros estadísticos, como análisis longitudinales y modelos de simulación y pronóstico, la validación y el procesamiento son muy estrictos. Las reglas para definir y verificar los objetos deben adaptarse a las necesidades de estos usos. Además,

el registro longitudinal debe complementarse con diversas variables importadas y variables derivadas nuevas que muchas veces requieren un alto grado de procesamiento.

El procesamiento relacionado con los usos avanzados se debe documentar con detalle. Esto también se aplica al procesamiento de registros para proyectos individuales de investigación. Esta documentación es necesaria para examinar la metodología y presentar ideas sobre futuras tareas y otros productos.

CAPÍTULO 7

Métodos de estimación

Luego de efectuar el procesamiento descrito en los capítulos 5 y 6, la matriz o matrices de datos del registro están listas para su uso. El siguiente paso consiste en utilizar la matriz de datos para elaborar tablas estadísticas que resulten relevantes según los objetivos de investigación de que se trate. En éste y los dos capítulos siguientes, describiremos los métodos de estimación que se aplican o podrían aplicarse para hacer estimaciones y tablas.

Abordaremos algunos problemas relacionados con la calidad y recomendaremos soluciones basadas en ciertos métodos de estimación. Algunos de estos métodos parten del principio de que los ponderadores también se usan en las estadísticas basadas en registros, de manera similar a las encuestas por muestreo. Estos ponderadores son un tipo especial de variables derivadas.

Cuando se usa la matriz de datos para crear tablas estadísticas, las celdas de las tablas contendrán frecuencias, sumas u otras medidas estadísticas. Cuando se usan ponderadores para fines de estimación, se calculan frecuencias o sumas ponderadas. En este capítulo presentamos una introducción general a los métodos de estimación aplicados en las encuestas basadas en registros.

En el capítulo 8, describiremos métodos de estimación útiles para manejar problemas con *valores perdidos*, *sobrecobertura* y *cambios de nivel en las series de tiempo*. Normalmente no se presta atención a estos aspectos de calidad cuando se producen estadísticas basadas en registros, pero los métodos presentados pueden usarse para compensar las fuentes de errores y, por ende, reducir los errores. Los métodos de estimación que veremos en ese capítulo se basan en ponderadores, calibración de ponderadores e imputaciones.

En el capítulo 9, veremos métodos de estimación que pueden ser de utilidad para manejar los problemas de calidad que surgen al integrar datos de diferentes registros. Esta integración puede dar origen a *variables de valores múltiples*, tipos especiales de errores e inconsistencias. Las variables de valores múltiples (mencionadas en la sección 3.3.2) son comunes en los sistemas de registros, en los que se integran datos de diferentes clases de objetos.

Dada la interacción de los registros de un sistema, los valores perdidos y otros problemas de calidad que haya en un registro también afectarán a otros registros que importen datos de ahí. Incluso el método de ajuste por no respuesta para un registro afecta a otros registros del sistema. Por ende, los métodos que proponemos deben funcionar dentro de todo el sistema para que las estadísticas de los diferentes registros sean consistentes.

Los métodos planteados en los capítulos 7, 8 y 9 se ilustran con ejemplos, que se han simplificado en aras de aclarar los principios generales. Algunos ejemplos se basan en datos reales de la Oficina de Estadística de Suecia y otros en pequeños conjuntos de datos ficticios.

7.1 ESTIMACIÓN EN ENCUESTAS POR MUESTREO Y ENCUESTAS BASADAS EN REGISTROS

Aunque el término *estimación* se aplica normalmente a las encuestas por muestreo, también debe usarse en las estadísticas basadas en registros. Asimismo, es importante distinguir entre los valores reales en la población objetivo y las estimaciones producidas con el registro.

7.1.1 Métodos de estimación basados en ponderadores

Todas las fórmulas siguientes se refieren a una celda en particular de una tabla, de modo que la notación está simplificada por esta restricción. En una celda de una tabla (sombreada en la tabla siguiente) hay R observaciones del registro y queremos estimar el total Y de la celda en la población de registro. En algunos casos, este total de la celda debe ajustarse debido a problemas de calidad.

		$y_1, y_2, \ldots y_R$		

Por cada celda de una tabla, se calculan *sumas* para variables cuantitativas, como salario o facturación usando las fórmulas (2) o (3) siguientes. Si se desea calcular las *frecuencias*, o número de observaciones, en las celdas, se usa la variable $y = 1$ en las fórmulas para todas las observaciones.

Con las *encuestas por muestreo*, se hacen estimaciones usando la fórmula (1) siguiente. Los ponderadores obtenidos en el diseño, d_i, dependen de cómo se ha diseñado la muestra o asignado en diferentes estratos. Los ponderadores g_i de la fórmula (1) se basan en las variables auxiliares de los registros estadísticos y sirven para minimizar el error de muestreo y los errores causados por la no respuesta. Deville y Särndal (1992) introducen este método de estimación.

$$\hat{Y} = \sum_{i=1}^{r} d_i g_i y_i = \sum_{i=1}^{r} w_i y_i \qquad \text{donde } r \text{ es el número de objetos en la } \textit{muestra} \text{ que respondieron en una celda en particular} \qquad (1)$$

Los ponderadores d_i son los ponderadores originales antes de la calibración y los ponderadores $d_i g_i = w_i$ son los ponderadores *después* de la calibración usando información sobre los totales de registro de algunas variables auxiliares de los registros estadísticos. Con la fórmula (1) se calculan las sumas y frecuencias ponderadas. Cuando se calculan los valores medios, las sumas ponderadas se dividen entre las frecuencias ponderadas.

En el caso de *censos* para los que la Oficina de Estadística de Suecia capta los datos, la no respuesta ocurre de la misma manera que en las encuestas por muestreo. Por ejemplo, usando los métodos descritos en Särndal y Lundström (2005), también es posible calcular los ponderadores g_i para reducir los errores causados por la no respuesta.

7.1.2 Métodos de estimación para las encuestas basadas en registros

En esta sección, analizaremos dos preguntas:

– ¿Hay métodos de estimación para las encuestas basadas en registros?

– ¿Se puede obtener una buena calidad con métodos de estimación en los que se usan ponderadores?

¿Hay métodos de estimación para las encuestas basadas en registros?

En la actualidad no se usan métodos especiales para producir estadísticas basadas en registros, sino que se hacen cálculos y sumas de la manera más simple posible.

$$\hat{Y} = \sum_{i=1}^{R} y_i \qquad \text{donde } R \text{ es el número de objetos en el } \textit{registro} \text{ en una celda en particular de la tabla} \tag{2}$$

Interpretamos estos cálculos aparentemente simples como estimaciones, cuyos valores dependen de los métodos usados cuando se creó el registro. El trabajo realizado para crear un registro estadístico se describe en los capítulos 5 y 6. Si este trabajo se realiza de distintas maneras, habrá diferente valores numéricos en las estadísticas que se producen con base en el registro. Elegir la metodología para crear un registro también significa elegir una metodología de estimación.

En las encuestas por muestreo, esta metodología se centra en *cómo* llevar a cabo la suma, es decir, cómo se decidirán los ponderadores d_i y g_i. En cambio, el trabajo metodológico para estadísticas basadas en registros se orienta a *qué* se debe sumar, es decir, cómo se formarán las variables del registro con los datos disponibles. La manera en que se crea un registro estadístico determina qué estimaciones se harán con el registro. De modo que también hay métodos de estimación dentro de la estadística basada en registros.

En este sentido, mientras que en los capítulos 5 y 6 nos referimos a métodos de estimación que consideramos *fundamentales* para las encuestas basadas en registros, en los capítulos 8 y 9 presentamos los denominados métodos de estimación *complementarios*.

¿Se puede obtener una buena calidad con métodos de estimación en los que se usan ponderadores?

Además de los métodos de estimación seleccionados por la manera en que se crea un registro, en los capítulos 8 y 9 introduciremos los ponderadores w_i para solucionar algunos problemas de calidad. Los ponderadores, que se calculan de distintas maneras dependiendo del problema, permiten corregir diferentes tipos de errores, por ejemplo, que las estimaciones de los registros estén en un nivel incorrecto.

En las encuestas basadas en registros, se usan los ponderadores $d_i = 1$ para objetos sin valores perdidos y $d_i = 0$ para objetos con valores perdidos. En este caso, las estimaciones se hacen aplicando la fórmula (3):

$$\hat{Y} = \sum_{i=1}^{R} d_i g_i y_i = \sum_{i=1}^{R} w_i y_i \qquad \text{donde } R \text{ es el número de objetos del } \textit{registro} \text{ en una celda en particular} \tag{3}$$

De acuerdo con los métodos tradicionales, todo $w_i = 1$, pero en los capítulos 8 y 9 se usarán otros ponderadores. Los tipos de errores que analizaremos en estos capítulos incluyen: errores derivados de la eliminación de información en variables de valores múltiples, errores derivados de una subcobertura, errores derivados de la no respuesta parcial o valores perdidos y errores ocasionados por cambios de nivel en las series de tiempo. La metodología podría usarse para más tipos de errores.

7.2 ENCUESTAS BASADAS EN REGISTROS: MÉTODOS DE ESTIMACIÓN FUNDAMENTALES

Como vimos en el capítulo 5, la creación de un registro estadístico se puede dividir en las siguientes fases. En todas ellas, el trabajo estadístico influye en las estimaciones que se

obtendrán con el registro. Las influencias de las fases 1 y 3 son indirectas, por lo que no las abordaremos aquí.

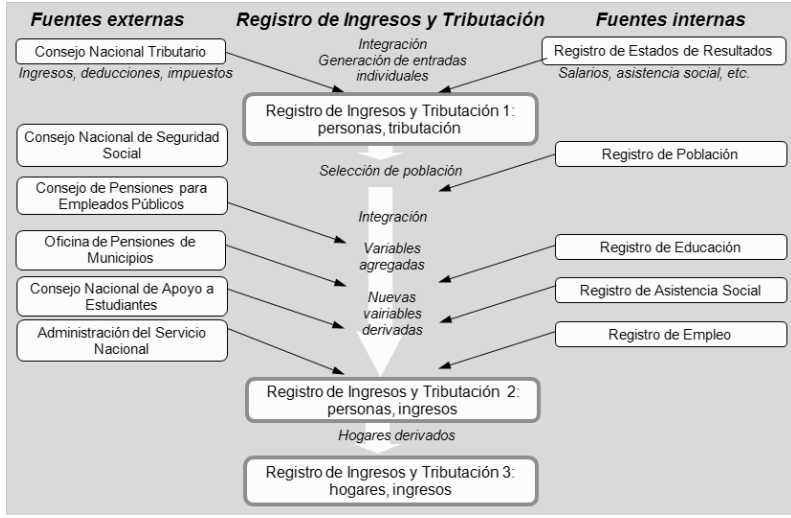
1. (Determinación de los objetivos de investigación).
2. Inventario: ¿cuáles son las fuentes disponibles cuando se crea un registro?
3. (Planeación).
4. Contactos con proveedores de datos y recepción de datos administrativos.
5. Integración, que se puede dividir en tres partes:
   a. ¿Cómo se deben integrar las fuentes que ya existen para que el registro contenga el conjunto de objetos necesario, es decir, la población de registro?
   b. ¿Qué procesamiento debe llevarse a cabo para verificar y corregir las definiciones de los objetos? Se forman los objetos derivados en el nuevo registro.
   c. Las variables de las fuentes administrativas se verifican y validan. ¿Cuál es el procesamiento para crear las variables en el registro? Se forman las variables derivadas en el nuevo registro.

Supongamos que queremos estimar el ingreso disponible equiparado (ingreso disponible promedio por unidad de consumo) de los hogares en diferentes regiones durante cierto año. En el siguiente gráfico (similar al gráfico 1.4 del capítulo 1) se ilustra el trabajo que se realiza para crear el Registro de Ingresos y Tributación.

En el gráfico, se muestran seis fuentes externas, pero en realidad se usan alrededor de 30 registros administrativos. Tanto el trabajo de *inventario* (fase 2), necesario para identificar estas fuentes, como el trabajo de *comunicación con cada autoridad administrativa* (fase 4), para que el equipo responsable del Registro de Ingresos y Tributación entienda y use correctamente cada fuente administrativa, tienen una repercusión muy importante en las estimaciones finales.

El *trabajo de validación* (fases 4 y 5, en particular 5c), en el que cada fuente se valida y corrige, y el trabajo final de validación de la consistencia de todas las fuentes en conjunto también son de gran relevancia para las estimaciones que se producirán con la versión final del Registro de Ingresos y Tributación.

Gráfico 7.1 Fuentes de datos y procesamiento de registros para el Registro de Ingresos y Tributación



La manera en que se *define* y crea la *población de registro* (fase 5a) es fundamental para las estimaciones del ingreso. Si la población se define como una población por año calendario, la suma del ingreso será mayor que si la población se define al final del año. También influirá en las estimaciones regionales si la población por región se define según el lugar donde el Consejo Tributario tiene registradas administrativamente a las personas físicas o si se usan los domicilios reales.

La unidad hogar en un sistema de registros es un tipo de objeto que se deriva con información administrativa. La manera en que se *definen y crean los hogares* (fase 5b) en el Registro de Ingresos y Tributación constituye una parte esencial de la estimación.

Por último, las variables derivadas se pueden crear de diferentes maneras. La variable *ingreso disponible equiparado* es una variable derivada y la manera en que se crea (fase 5c) también es una parte muy importante del proceso de estimación.

7.3 USO DE PONDERADORES EN ENCUESTAS BASADAS EN REGISTROS

En los métodos de estimación fundamentales presentados en los capítulos 5 y 6 no se usan ponderadores; en cambio, en los métodos de estimación complementarios de los capítulos 8 y 9 se utilizan diferentes tipos de ponderadores para reducir diferentes clases de errores o problemas de calidad.

De una matriz de datos a tablas estadísticas

¿Cómo se pueden usar los ponderadores en un registro para hacer estimaciones? Esto lo vemos en el registro salarial ficticio del gráfico siguiente. El registro contiene datos sobre el salario mensual durante un mes específico y datos sobre el horario de trabajo *(jornada)*, donde 1 es tiempo completo.

El salario real (en coronas suecas, considerando que ocho coronas equivalen aproximadamente a un dólar estadounidense o un euro) se recalcula para obtener un salario de tiempo completo usando la variable *jornada*. La ocupación de la persona se indica con el código ocupacional (Clasificación Internacional Uniforme de Ocupaciones, CIUO). La variable ocupacional se puede agrupar por nivel de competencia, donde 1 es el más bajo y 5 se refiere a directivos. Las variables *edad* y *salario de tiempo completo* también se pueden dividir en clases. Para corregir cualquier error, por ejemplo, la no respuesta, se han calculado los ponderadores w_i.

Gráfico 7.2 Registro salarial que incluye ponderadores en sus observaciones

Persona	Sexo	Edad	CIUO	Nivel	Salario	Jornada	Salario completo	Clase salarial	w_i	$w_i \cdot$ Salario$_i$	$w_i \cdot$ Jornada$_i$	$w_i \cdot$ Salario completo$_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
NIP1	F	50–54	4 190	2	14 850	1.00	14 850	14–14.9	1.028	15 271.4	1.028	15 271.4
NIP2	F	40–44	2 330	4	16 630	0.95	17 505	17–17.9	1.031	17 147.5	0.980	18 049.8
NIP3	M	50–54	2 492	4	17 807	1.00	17 807	17–17.9	1.083	19 285.5	1.083	19 285.5
NIP4	F	40–44	2 330	4	1 485	0.09	16 500	16–16.9	1.031	1 531.2	0.093	17 013.5
NIP5	F	40–44	5 133	2	6 497	0.50	12 994	12–12.9	1.031	6 699.2	0.516	13 398.4
NIP6	F	40–44	5 131	2	14 102	1.00	14 102	14–14.9	1.031	14 540.9	1.031	14 540.9
NIP7	M	50–54	5 131	2	858	0.06	14 300	14–14.9	1.083	929.2	0.065	15 487.3
...	...	...	...	...	...	...	...	...	...	...	...	...

El registro contiene las columnas (1) a (10) y, cuando se efectúan las estimaciones, se forman temporalmente las columnas (11) a (13). La tabla fuente que se presenta a

continuación se genera sumando las variables de respuesta en las columnas (10) a (13) del registro para todas las combinaciones de las variables de expansión sexo, edad, CIUO (que incluye también nivel de competencia) y clase salarial.

La tabla completa que aparece en el gráfico 7.3 consta de alrededor de 2,200 renglones. Las encuestas basadas en registros son adecuadas para formar tablas fuente grandes y detalladas que, a su vez, pueden servir si se desea crear tablas legibles para distintos propósitos.

Gráfico 7.3 Una tabla fuente se puede usar para formar muchas otras tablas

Sexo	Edad	CIUO	Nivel	Clase salarial	$\sum w_i$	$\sum w_i \cdot$ Salario$_i$	$\sum w_i \cdot$ Jornada$_i$	$\sum w_i \cdot$ Salario completo$_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
F	17–24	2330	4	12–12.9	42.52	429 170	34.55	526 165
F	17–24	2330	4	13–13.9	95.67	1 293 410	95.35	1 297 704
F	17–24	2330	4	14–14.9	42.52	201 399	14.14	622 852
F	17–24	2330	4	15–15.9	10.63	159 444	10.63	159 444
F	17–24	2330	4	16–15.9	53.15	163 111	9.89	876 942
...	...	...	...	...	...	...	...	...
M	60–64	8320	2	13–13.9	21.24	290 107	21.23	290 107
M	60–64	8320	2	14–14.9	10.62	149 300	10.62	149 300
M	60–64	9140	1	12–12.9	10.62	136 422	10.62	136 422
M	60–64	9140	1	13–13.9	10.62	71 348	5.31	142 697
M	60–64	9140	1	14–14.9	21.24	308 040	21.23	308 040

Al seguir agregando esta tabla fuente de diferentes maneras, se pueden formar más tablas para diferentes propósitos. La variable *salario* se usa como variable de expansión (columna 5) y como variable de respuesta (columnas 7 y 9) en la tabla fuente anterior. En una tabla estadística real, el salario sólo se usa en una de estas funciones.

Recuperación de las tablas de frecuencia: distribución salarial para diferentes dominios de estudio

Al sumar la columna (6) del gráfico 7.3, que contiene el número ponderado de personas, para diferentes combinaciones de nivel de competencia, sexo y clase salarial, se produce el gráfico 7.4. Las frecuencias absolutas del gráfico A se recalculan para obtener las frecuencias relativas de la tabla del gráfico B e ilustrar la relación entre sexo y salario, en primer lugar, y entre nivel de competencia y salario, en segundo lugar.

Gráfico 7.4 Distribución salarial por nivel y sexo

A. Número de personas

Nivel:	**1**		**2**		**3**		**4**		**5**	
Salario	M	H	M	H	M	H	M	H	M	H
9–11.9	615	107	1 823	484	83	32	31			
12–12.9	1 138	108	2 806	434	199	32	73			
13–13.9	2 220	381	10 382	1 686	239	130	397	174		
14–14.9	560	162	9 675	968	900	347	831	194		
15–15.9	114	54	4 246	565	1 719	533	911	228	21	
16–15.9	21	44	1 709	651	1 758	467	1 293	454	10	
17–17.9			1 389	520	1 054	468	1 675	576	124	
18–18.9	10	11	765	251	786	271	1 729	721	114	11
19–19.9			196	122	487	229	1 076	644	21	11
20–20.9			73	43	289	110	1 492	882	31	21
21–22.9			21	22	237	66	550	567	62	44
23–25.9				11		22	238	412	238	250
26–29.9					10	11	114	205	186	163
30–34.9							52	151	10	99
35–39.9							155	230		44
40–125							145	492		33
Total	4 677	869	33 084	5 758	7 762	2 717	10 763	5 930	817	675

B. Porcentaje

1		...	**5**	
M	H		M	H
13.2	12.3	...	**0.0**	0.0
24.3	12.5	...	**0.0**	0.0
47.5	43.8	...	**0.0**	0.0
12.0	**18.7**	...	**0.0**	0.0
2.4	**6.3**	...	**2.5**	0.0
0.4	**5.1**	...	**1.3**	0.0
0.0	**0.0**	...	**15.2**	0.0
0.2	**1.3**	...	**13.9**	1.6
0.0	**0.0**	...	**2.5**	1.6
0.0	**0.0**	...	**3.8**	3.2
0.0	**0.0**	...	**7.6**	6.4
0.0	**0.0**	...	29.1	**37.1**
0.0	**0.0**	...	22.7	**24.2**
0.0	**0.0**	...	1.3	**14.6**
0.0	**0.0**	...	0.0	**6.5**
0.0	**0.0**	...	0.0	**4.9**
100.0	100.0	...	100.0	100.0

Recuperación de tablas con medias y cocientes para diferentes ámbitos de estudio

Al sumar la variables de respuesta en las columnas (6) a (9) de la tabla fuente presentada en el gráfico 7.3 para diferentes combinaciones de las variables de expansión en las columnas (1) a (4), se forman tablas con totales que pueden servir para formar valores medios y cocientes en diferentes ámbitos de estudio.

Gráfico 7.5 Tabla base para varias otras tablas

Sexo	Nivel	$\sum w_i$	$\sum w_i \cdot \text{salario}_i$	$\sum w_i \cdot \text{jornada}_i$	$\sum w_i \cdot \text{salario completo}_i$	(6)/(3)	(4)/(5)	(5)/(3)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
F	1	4 677	40 635 041	3 061	61 701 517	13 191	13 274	0.65
F	2	33 084	360 320 838	25 085	474 211 547	14 334	14 364	0.76
F	3	7 762	106 558 725	6 416	129 063 365	16 627	16 607	0.83
F	4	10 763	175 129 111	9 313	202 168 595	18 784	18 805	0.87
F	5	817	17 415 674	771	18 333 690	22 431	22 583	0.94
M	1	869	10 905 658	797	11 795 151	13 579	13 690	0.92
M	2	5 758	73 156 416	4 898	85 083 524	14 777	14 936	0.85
M	3	2 717	42 580 548	2 532	45 632 570	16 796	16 815	0.93
M	4	5 930	125 232 247	5 430	136 222 043	22 971	23 063	0.92
M	5	675	18 050 648	655	18 594 809	27 531	27 564	0.97

Gráfico 7.6

Salario medio por sexo y nivel de competencia

Nivel	M	H
1	13 274	13 690
2	14 364	14 936
3	16 607	16 815
4	18 805	23 063
5	22 583	27 564

La tabla anterior se forma usando la columna (8) de la tabla del gráfico 7.5.

En el gráfico 7.5, la columna (3) de la tabla contiene el número ponderado de personas dentro de cada ámbito de estudio, la columna (4) los salarios reales ponderados, la columna (5) los empleos de tiempo completo ponderados, la columna (6) los salarios de tiempo completo ponderados, la columna (7) el salario de tiempo completo promedio por persona, la columna (8) el salario promedio por empleo de tiempo completo y la columna (9) la jornada de trabajo promedio por persona.

La tabla del gráfico 7.5 puede usarse para crear otras tablas, que deberán tener el formato de la tabla del gráfico 7.6, donde en primer lugar se puede comparar el salario medio para hombres y mujeres.

La tabla del gráfico 7.7 muestra partes de una tabla en la que los salarios promedio de mujeres y hombres se comparan dentro de la misma categoría de edad y ocupación. La tabla completa consta de alrededor de 800 renglones y, por lo tanto, es apropiada para resumir su contenido calculando los salarios medios estandarizados. Las encuestas basadas en registros son adecuadas para análisis detallados de tablas que a menudo deberán complementarse con el cálculo de medias estandarizadas a partir de ponderadores estándares.

Las mujeres y los hombres presentan una distribución distinta por edad y ocupación, lo que explica la mayor parte de la diferencia salarial: 15 680−18 860. Si mujeres y hombres tuvieran la misma distribución por edad y ocupación, de acuerdo con los ponderadores estándares de la columna (8), la diferencia salarial sólo sería de 16 256−16 505 de acuerdo con las dos últimas columnas.

Gráfico 7.7 Cálculo de salarios medios estandarizados

Edad	CIUO	Salario promedio, mujeres	Salario promedio, hombres	Número de mujeres	Número de hombres	Número total	Ponderación estándar	M	H
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) = (7) / 5 688	(3) · (8)	(4) · (8)
17–24	2330	13 660	14 100	276	75	351	0.0062	84.41	87.13
…	…	…	…	…	…	…	…	…	…
60–64	7130	13 826	13 900	10	63	74	0.0013	18.09	18.19
Total		**15 680**	**18 860**	4 523	1 165	5 688	1.0000	**16 256**	**16 505**

7.4 ESTIMACIÓN MEDIANTE PONDERADORES Y REGISTROS POR AÑO CALENDARIO

En esta sección presentaremos ejemplos simples en los que se deben usar ponderadores a fin de producir estimaciones para estadísticas basadas en registros. El registro por año calendario es la versión del registro que contiene todos los objetos que han existido en cualquier momento durante un año específico. En un registro por año calendario, pueden existir objetos durante diferentes periodos. La mayoría de los objetos no causan problemas porque existen a lo largo de todo el año, pero hay otros que nacen o entran en el registro en ciertos momentos y otros más que desaparecen o mueren durante el año calendario. Esto da origen a problemas de estimación que se pueden resolver mediante ponderadores. El tiempo se puede usar como una *variable generadora de ponderadores*; con estos ponderadores se pueden producir estimaciones correctas para los registros por año calendario.

Población promedio

La población promedio de un municipio se puede estimar de la siguiente manera, donde calculamos el día de nacimiento o llegada al municipio como un día completo y el día en que la persona se mudó o murió como ningún día en el municipio.

Gráfico 7.8 Registro por año calendario para la población en un municipio (pequeño)

Perso-na	Existía el 1/1 2005	Llegó durante 2005 aaaammdd	Desapareció durante 2005 aaaammdd	Existía el 31/12 2005	Ponderador = tiempo en el municipio, años
NIP1	Sí	-	20050517	No	136/365 = 0.37
NIP2	Sí	-	-	Sí	365/365 = 1.00
NIP3	No	20050315	20050925	No	194/365 = 0.53
NIP4	No	20050606	-	Sí	209/365 = 0.57
Total	**2**			**2**	**2.47**

La manera tradicional de calcular la población promedio para 2005 consiste en obtener el valor promedio de la población al 1 de enero de 2005 (2 personas) y la población al 31 de diciembre de 2005 (también 2 personas). Un cálculo más específico, en el que el tiempo en el municipio se usa como ponderador, arroja como resultado que la población promedio durante 2005 fue de 2.47 en vez de la medida tradicional de 2.

Variables de flujo y de stock

Los datos del ejemplo siguiente se refieren a empresas de una región en particular durante 2004. Ciertas empresas empezaron a existir o dejaron de existir en diferentes momentos del año, de modo que en este caso el tiempo también puede usarse como variable generadora de ponderadores.

Las variables de flujo y de *stock* deben tratarse de manera distinta. Las variables de flujo, como el valor agregado de una empresa, sólo se relacionan con los valores durante el periodo del año en que la empresa estuvo activa y, por ende, no es necesario ponderarlas. En cambio, sí debe ponderarse una variable de *stock* que muestra el nivel en determinado momento, como el número de empleados. El valor agregado total en la región durante 2004 ascendió a 83 millones de coronas, mientras que el número promedio de empleados fue de 112.5. La productividad se calcula como 83/112.5 = 0.738 millones de coronas anuales por empleado.

Gráfico 7.9 Registro por año calendario correspondiente a 2004 para empresas en una región (pequeña) en particular

Identidad empresarial	Existía al 1/1	Llegó	Dejó de existir	Existía al 31/12	Ponde-rador	Valor agregado	Número de empleados	Ponderador • Número de empleados
UE1	Sí	-	20040630	No	0.50	10	30	0.50 • 30 = 15.0
UE2	Sí	-	-	Sí	1.00	42	45	1.00 • 45 = 45.0
UE3	No	20040401		Sí	0.75	31	70	0.75 • 70 = 52.5
Total					**2.25**	**83**		**112.5**

7.5 CALIBRACIÓN DE LOS PONDERADORES EN ENCUESTAS BASADAS EN REGISTROS

En la sección 7.1.2 señalamos que una manera general de estimar estadísticas basadas en registros consiste en usar la fórmula (3) siguiente:

$$\hat{Y} = \sum_{i=1}^{R} d_i g_i y_i = \sum_{i=1}^{R} w_i y_i \qquad \text{donde } R \text{ es el número de objetos del registro en una celda en particular} \qquad (3)$$

En esta sección ilustramos cómo se pueden calibrar los ponderadores mediante un ejemplo basado en el registro que aparece en el gráfico 7.10. De las 19 observaciones que hay en el registro, dos (la 6 y la 15) tienen valores perdidos. Cuatro personas no están empleadas y, por lo tanto, aunque no tienen clasificación industrial, no son valores perdidos.

Gráfico 7.10 Registro sobre personas de dos regiones pequeñas

(1) NIP	(2) Sexo	(3) Distrito	(4) Empleado	(5) Rama industrial	(6) Nivel de escolaridad	(7) d_i	x_{1i} Sexo=F	x_{2i} Sexo=M	x_{3i} Distrito=1	x_{4i} Empleado=1	w_i
1	F	1	0	nulo	Bajo	1	1	0	1	0	0.98276
2	M	1	1	A	Bajo	1	0	1	1	1	1.15517
3	F	1	1	A	Bajo	1	1	0	1	1	1.13793
4	M	1	1	A	Medio	1	0	1	1	1	1.15517
5	F	1	1	A	Medio	1	1	0	1	1	1.13793
6	M	1	1	***Perdido***	Bajo	0	0	1	1	1	0.00000
7	F	1	1	D	Medio	1	1	0	1	1	1.13793
8	M	1	1	D	Alto	1	0	1	1	1	1.15517
9	F	1	1	D	Medio	1	1	0	1	1	1.13793
10	M	1	0	nulo	Medio	1	0	1	1	0	1.00000
11	F	2	0	nulo	Bajo	1	1	0	0	0	1.00000
12	M	2	1	D	Bajo	1	0	1	0	1	1.17241
13	F	2	1	D	Bajo	1	1	0	0	1	1.15517
14	M	2	1	D	Medio	1	0	1	0	1	1.17241
15	F	2	1	D	***Perdido***	0	1	0	0	1	0.00000
16	M	2	1	A	Bajo	1	0	1	0	1	1.17241
17	F	2	1	A	Medio	1	1	0	0	1	1.15517
18	F	2	1	A	Medio	1	1	0	0	1	1.15517
19	M	2	0	nulo	Medio	1	0	1	0	0	1.01724

Si queremos estimar una tabla de frecuencias que describa el nivel de escolaridad por rama industrial con este registro, los valores perdidos afectarán las estimaciones. La tabla del

gráfico 7.11 se basa en las columnas sombreadas del gráfico 7.10 y en sumas simples con los ponderadores d_i.

Gráfico 7.11 Personas por nivel de escolaridad y rama industrial

	Rama industrial A Núm. de personas	Rama industrial D Núm. de personas	Rama industrial A Porcentaje	Rama industrial D Porcentaje
Nivel de escolaridad alto	0	1	0.0%	16.7%
Nivel de escolaridad medio	4	3	57.1%	50.0%
Nivel de escolaridad bajo	3	2	42.9%	33.3%
Todos	7	6	100.0%	100.0%

Las variables de las columnas (2), (3) y (4) no tienen valores perdidos y se pueden usar para calibrar los ponderadores d_i, de modo que las estimaciones que usan los ponderadores calibrados w_i se ajustarán para los valores perdidos en las columnas (5) y (6).

Las sumas o frecuencias basadas en las variables sin valores perdidos se pueden usar como condiciones para la calibración. Hay muchas maneras de elegir estas condiciones y para cada elección tenemos ponderadores calibrados que pueden diferir. En este ejemplo usaremos cuatro condiciones: número correcto de mujeres = 10, de hombres = 9, de personas en el distrito 1 = 10 y de personas empleadas = 15.

Si estas cuatro frecuencias se estiman a partir del conjunto de observaciones con valores perdidos, deberán usarse los ponderadores d_i y, a causa de los valores perdidos, las estimaciones de las mismas estadísticas serán erróneas: el número de mujeres = 9 (error = –1), de hombres = 8 (error = –1), de personas en el distrito 1 = 9 (error = –1) y de empleados = 13 (error = –2).

La idea de la calibración es ajustar los ponderadores d_i de manera que estas cuatro estimaciones no presenten errores. Cualquier otra estimación también se ajustará de la misma forma. Usando los nuevos ponderadores, se pueden producir estimaciones consistentes ajustadas para los valores perdidos del registro.

Las primeras siete columnas del gráfico 7.10 muestran el registro original, mientras que las columnas $x_{1i} - x_{4i}$ contienen información que se usará en la calibración. En los cálculos, se usan vectores $\boldsymbol{x}_i'$, uno por renglón. Para *i=1*, como para *PIN1*, $\boldsymbol{x}_1' =$ (1 0 1 0).

Ahora las sumas se refieren a todas las observaciones incluidas en el registro, no sólo a una celda como en las fórmulas anteriores (1)–(3). La última columna del gráfico 7.10 muestra los ponderadores ajustados w_i, calculados en tres etapas:

1. Se calculan $\boldsymbol{T} = \Sigma d_i \boldsymbol{x}_i \boldsymbol{x}'_i$ y $\boldsymbol{T}^{-1}$, donde todo $d_i = 1$ (valores perdidos, $d_i = 0$) y $i = 1, 2, \ldots, 19$
   $\boldsymbol{T}$ es una matriz con cuadrados y productos totales. Aquí el caso de la matriz de 4 × 4:

$$\boldsymbol{T} = \begin{bmatrix} \Sigma d_i x^2_{1i} & \Sigma d_i x_{1i} x_{2i} & \Sigma d_i x_{1i} x_{3i} & \Sigma d_i x_{1i} x_{4i} \\ \Sigma d_i x_{2i} x_{1i} & \Sigma d_i x^2_{2i} & \Sigma d_i x_{2i} x_{3i} & \Sigma d_i x_{2i} x_{4i} \\ \Sigma d_i x_{3i} x_{1i} & \Sigma d_i x_{3i} x_{2i} & \Sigma d_i x^2_{3i} & \Sigma d_i x_{3i} x_{4i} \\ \Sigma d_i x_{4i} x_{1i} & \Sigma d_i x_{4i} x_{2i} & \Sigma d_i x_{4i} x_{3i} & \Sigma d_i x^2_{4i} \end{bmatrix}$$

2. Se calcula el vector $\boldsymbol{\lambda}$: $\boldsymbol{\lambda} = \boldsymbol{T}^{-1} (\boldsymbol{t}_x - \Sigma d_i \boldsymbol{x}_i)$
   El vector $\boldsymbol{t}_x$ representa las 4 condiciones para el número de mujeres y hombres, personas en el distrito 1 y personas empleadas.
   El vector $\Sigma d_i \boldsymbol{x}_i$ es el número no ajustado correspondiente.

t_x	$\Sigma d_i x_i$	$t_x - \Sigma d_i x_i$
10	9	1
9	8	1
10	9	1
15	13	2

El vector t_x representa los valores correctos de las cuatro condiciones de calibración y el vector $\Sigma d_i x_i$ los valores erróneos basados en las observaciones con valores perdidos.

3. Los ponderadores ajustados se convierten en: $w_i = d_i\,(1 + x'_i \lambda)$. Estos ponderadores se usan para calcular los números y los totales ponderados.

A continuación ilustramos estas fórmulas, con los cálculos paso a paso.

1. Se calculan las matrices T y T^{-1}:

$$T = \begin{bmatrix} 9 & 0 & 5 & 7 \\ 0 & 8 & 4 & 6 \\ 5 & 4 & 9 & 7 \\ 7 & 6 & 7 & 13 \end{bmatrix}$$

$$T^{-1} = \begin{bmatrix} 0.375000 & 0.250000 & -0.125000 & -0.250000 \\ 0.250000 & 0.362069 & -0.112069 & -0.241379 \\ -0.125000 & -0.112069 & 0.237069 & -0.008621 \\ -0.250000 & -0.241379 & -0.008621 & 0.327586 \end{bmatrix}$$

2. Se calcula el vector λ:

$$\lambda = \begin{bmatrix} 0.375000 & 0.250000 & -0.125000 & -0.250000 \\ 0.250000 & 0.362069 & -0.112069 & -0.241379 \\ -0.125000 & -0.112069 & 0.237069 & -0.008621 \\ -0.250000 & -0.241379 & -0.008621 & 0.327586 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} 0.000000 \\ 0.017241 \\ -0.017241 \\ 0.155172 \end{bmatrix}$$

3. Los ponderadores ajustados se convierten en: $w_i = d_i\,(1 + x'_i \lambda)$

   Para la primera persona en el registro, $i=1$ y $x_1' = (1\ 0\ 1\ 0)$

4.

$$x'_1 \lambda = [1\ 0\ 1\ 0] \bullet \begin{bmatrix} 0.000000 \\ 0.017241 \\ -0.017241 \\ 0.155172 \end{bmatrix} = -0.017241$$

   El ponderador ajustado para la persona 1 se convierte en:
   $w_1 = 1 \bullet (1 - 0.017241) = 0.982759$

A partir de los ponderadores ajustados se estiman las frecuencias ponderadas de la tabla siguiente. Las frecuencias relativas resultan ser casi las mismas que en el gráfico 7.11, pero ahora el número de personas asciende a 15 (8.1 + 6.9) en lugar de 13.

Gráfico 7.12 Personas por nivel de escolaridad y rama industrial, ajustadas por valores perdidos

	Rama industrial A, Núm. de personas ponderado	Rama industrial D, Núm. de personas ponderado	Rama industrial A Porcentaje	Rama industrial D Porcentaje
Nivel de escolaridad alto	0.0	1.2	0.0%	16.7%
Nivel de escolaridad medio	4.6	3.4	57.0%	49.7%
Nivel de escolaridad bajo	3.5	2.3	43.0%	33.6%
Todos	8.1	6.9	100.0%	100.0%

CAPÍTULO 8

Calibración e imputación

En este capítulo abordaremos tres temas: valores perdidos, sobrecobertura y cambios de nivel en las series de tiempo. En los tres casos, los ponderadores y su calibración se pueden usar como métodos de estimación complementarios, y la imputación puede ser utilizada para ajustar los datos por valores perdidos.

8.1 EL PROBLEMA DE LA NO RESPUESTA

Hoy son comunes los ajustes por no respuesta en las encuestas por muestreo, aunque no en las encuestas basadas en registros. Las tablas siguientes, que contienen dos ejemplos de la Oficina de Estadística de Suecia, ilustran dos formas de publicar estadísticas con no respuesta.

Encuesta sobre Fuerza Laboral 2001

Categoría de fuerza laboral	Miles	% de la población
Empleados	4 239	75.3
Desempleados	175	3.1
Fuera de la fuerza laboral	1 218	21.6
Población de 16 a 64 años	5 632	100.0

Nota: La tasa de no respuesta en la Encuesta sobre Fuerza Laboral es de aproximadamente 15%. Las estimaciones publicadas se han ajustado por no respuesta.

Registro de Educación 2001

Nivel de escolaridad	Miles	% de la población
Menos de 9 años	755	11.8
9 años obligatorios	939	14.7
Media superior: 2 años	1 747	27.4
Media superior: 3 años	1 142	17.9
Universidad: < 3 años	802	12.6
Universidad: ≥ 3 años	848	13.3
Posgrado	48	0.7
No especificado	**106**	**1.7**
Población de 16 a 74 años	6 386	100.0

Los valores perdidos en los registros se pueden manejar de diferentes maneras. Una posibilidad, bastante común, es publicar las tablas con una categoría, *no especificado*, y no hacer ningún ajuste si hay valores perdidos. Una segunda posibilidad consiste en usar ponderadores, calibrados para reducir los efectos de los valores perdidos. Una tercera posibilidad es imputar valores. Analizaremos estas tres opciones y, al final de esta sección, presentaremos algunas conclusiones sobre cómo manejar los valores perdidos en un registro. Asimismo, en las encuestas basadas en registros hay posibilidades de reducir la no respuesta (parcial), lo cual siempre deberá hacerse antes de aplicar los métodos de ajuste. Esta reducción se efectúa cuando se crea el registro y se deben utilizar todas las fuentes que sea posible para minimizar la no respuesta parcial.

Los métodos utilizados en encuestas por muestreo para hacer ajustes en caso de no respuesta se analizan en muchos libros y artículos; no obstante, los valores perdidos en los sistemas de registros y en las encuestas basadas en registros son un tema que hasta ahora no se ha investigado a fondo. La coordinación y consistencia entre los registros del sistema se suma

a los requisitos para los métodos de ajuste, un aspecto que normalmente no se aborda en la bibliografía sobre ajuste para encuestas por muestreo. Los métodos de ajuste para registros deben tomar como base esos métodos, pero es necesaria una mayor elaboración.

Nuestro propósito no es hacer un análisis general del ajuste en caso de no respuesta, sino sólo plantear algunos problemas específicos de los registros que forman parte de un sistema. Usaremos la versión más simple posible de cada método porque pensamos que la mayoría de la gente que trabaja con registros estadísticos no está acostumbrada a realizar ajustes por no respuesta. Särndal y Lundström (2005) también examinan los ajustes basados en los ponderadores y la calibración de los ponderadores, mientras que Eltinge *et al*. (2003) y Pannekoek y de Waal (2005) presentan una introducción y los aspectos generales de la imputación.

8.1.1 Valores perdidos en los registros

Hay dos formas de no respuesta en una encuesta con captación de datos propia: *no respuesta de un objeto* (no se recibe respuesta de ciertos objetos) y *no respuesta parcial* (no se responden ciertas preguntas, pero sí hay respuestas para otras preguntas). El concepto de no respuesta está formulado para encuestas con una captación de datos propia.

En el caso de los estudios basados en registros, resulta difícil diferenciar entre los conceptos de *no respuesta de un objeto* y *subcobertura*. El término para la no respuesta más adecuado en este tipo de encuestas es la *no respuesta parcial*, para indicar que ciertos valores están perdidos. En las *encuestas basadas en registros*, los valores perdidos pueden deberse a varias razones:

- Puede haber valores de variables perdidos para ciertos objetos a causa de fallas en el sistema administrativo.
- En estas encuestas, a menudo se crean registros a partir de *varios* registros fuente. Cuando se vinculan y emparejan diferentes registros fuente, puede haber no emparejamientos porque hay objetos perdidos en algunos de esos registros. Esto tiene como consecuencia valores perdidos para todas las variables importadas de ellos.
- En el proceso de validación (sección 6.3), es posible identificar valores perdidos o tomar la decisión de rechazar ciertos valores.

Cuando se trata de variables de un registro estadístico, siempre se debe notificar la magnitud de la no respuesta parcial. A continuación describiremos los métodos de ajuste por no respuesta. Si se imputan valores de variables, éstos se deben mostrar en variables especiales para que siempre quede claro cuáles son los valores medidos y cuáles los imputados.

Al crear un registro, se importan variables de diferentes registros fuente. Si algunas de estas variables importadas contienen no respuesta parcial, el nuevo registro incluirá un mayor número de no respuesta parcial que las fuentes. Hay ciertas variables cualitativas usadas en muchos registros en los que la no respuesta debe tratarse de manera consistente dados los estrictos requisitos de consistencia.

8.1.2 No haga ajustes, publique el "valor conocido"

En esta sección presentamos algunos ejemplos donde no se hicieron ajustes por la no respuesta parcial. Nos interesa mostrar los inconvenientes de esta manera de publicar estadísticas.

Comparación de áreas pequeñas

El Registro de Empleo sueco contiene datos acerca de personas con un empleo remunerado, con una clasificación industrial para el establecimiento donde la persona está empleada y su nivel de escolaridad. Estas variables se importan del Registro Empresarial y el Registro de Educación. Ambas fuentes contienen no respuesta parcial o valores perdidos.

De acuerdo con los datos del gráfico 8.1, la tasa de valores perdidos para nivel de escolaridad es de 1.7 % y para rama industrial de 79/5647 = 1.4%. De la población del Registro de Empleo con respecto a toda la población de 16 a 64 años, (79 + 93 – 2)/5647 = 3.0 % no tiene valores para clasificación industrial *o* nivel de escolaridad.

Gráfico 8.1
Población de 16 a 64 años por nivel de escolaridad y clasificación industrial, 2001

Miles de personas	Empleados en la rama industrial				No empleados	Población total	Rama industrial				No empleados	Población total
	A–F producción de bienes	G–K servicios privados	L–Q servicios públicos	No especificado			A–F %	G–K %	L–Q %	No especificado %	%	%
<9 años, 9 años obligatorios	273	273	141	17	611	1315	24.7	18.9	9.7	21.8	39.3	23.3
Media superior: 2 años	403	404	416	22	324	1570	36.6	27.9	28.5	28.5	20.8	27.8
Media superior: 3 años	229	369	217	15	252	1081	20.7	25.5	14.9	18.7	16.2	19.1
Universidad: < 3 años	107	188	273	11	177	757	9.7	13.0	18.7	13.6	11.4	13.4
Universidad: ≥ 3 años	86	204	410	12	119	830	7.8	14.1	28.0	15.4	7.7	14.7
No especificado	5	10	4	**2**	73	**93**	0.4	0.7	0.3	2.0	4.7	1.7
Población total de 16 a 64 años	1 103	1 448	1 462	**79**	1 556	5 647	100	100	100	100	100	100

La tabla anterior es un ejemplo de la manera en que la Oficina de Estadística de Suecia ha presentado tradicionalmente las estadísticas basadas en registros con valores perdidos. Esto se debe a varias razones: la tasa de valores perdidos se considera menor y se ha juzgado que sería demasiado complicado hacer ajustes. En nuestra opinión, la tasa de valores perdidos no es menor, es muy factible hacer ajustes y la Oficina de Estadística de Suecia tiene la responsabilidad de hacer los ajustes necesarios para valores perdidos.

En la anterior tabla de dos dimensiones, hay valores perdidos en ambas variables de expansión. Aunque el valor perdido total sólo ascienda a 3.0%, se obtiene una tabla difícil de interpretar porque los valores perdidos que contiene la tabla alteran los patrones de comportamiento de las variables. Haciendo los ajustes correspondientes, las tablas serán más fáciles de interpretar para los usuarios de las estadísticas.

Sin embargo, la magnitud de esta tasa de valores perdidos varía cuando se comparan *municipios* u otras categorías pequeñas. En la tabla siguiente aparecen los porcentajes máximos y mínimos de valores perdidos para los municipios del país.

Gráfico 8.2 Registro de Empleo 2000 y 2001: valores máximos y mínimos de la tasa de no respuesta parcial, población de 16 a 64 años en los municipios suecos

No respuesta parcial:	Mínimo en 2000	Máximo en 2000	Mínimo en 2001	Máximo en 2001
Nivel de escolaridad no especificado	0.3%	3.7%	0.4%	4.6%
Rama industrial desconocida	0.5%	6.0%	0.4%	3.7%
Nivel de escolaridad y rama industrial no especificados	1.0%	7.0%	1.0%	7.5%

Aunque la tasa de valores perdidos no sea tan alta a escala nacional, puede ser considerable en los municipios y también puede variar de un año a otro. Observamos esto en el gráfico

8.2, lo que significa que la no respuesta parcial dificulta las comparaciones entre municipios y otras categorías pequeñas. Muchos usuarios olvidarán la categoría "rama industrial desconocida" cuando comparen los diferentes municipios en lo relativo, por ejemplo, al tamaño del sector de servicios.

¿Cómo comparar un municipio con una tasa de valores perdidos del 1% con otro municipio con una tasa de valores perdidos del 7%? Para hacer esa comparación será necesario hacer ajustes por valores perdidos; de lo contrario, las estimaciones de los niveles para los distintos municipios no serán comparables. En el gráfico 8.3, hemos hecho ajustes para los valores perdidos mostrados en el gráfico 8.1. Las personas con "nivel de escolaridad no especificado" se distribuyeron proporcionalmente en los niveles de escolaridad de la misma columna y se hizo lo propio con las personas con "rama industrial desconocida" en las tres ramas industriales en el mismo renglón.

Gráfico 8.3 Población de 16 a 64 años de edad por nivel de escolaridad y clasificación industrial, 2001. Estimaciones ajustadas por valores perdidos

Miles de personas	Rama industrial			No empleados	Población total	Rama industrial			No empleados	Población total
	A–F producción de bienes	G–K servicios privados	L–Q servicios públicos			A–F	G–K	L–Q		
						%	%	%	%	%
<9 años, 9 años obligatorios	281	282	145	641	1 337	25.0	19.1	9.8	41.2	23.7
Media superior: 2 años	413	415	425	340	1 597	35.7	28.1	28.6	21.8	28.3
Media superior: 3 años	234	378	222	264	1 100	20.8	25.6	14.9	17.0	19.5
Universidad: < 3 años	110	193	279	186	770	9.8	13.1	18.8	12.0	13.6
Universidad: ≥ 3 años	88	209	418	125	844	7.8	14.1	28.1	8.0	14.9
Población total de 16 a 64 años	1 125	1 476	1 490	1 556	5 647	100.0	100.0	100.0	100.0	100.0

La tabla ajustada para valores perdidos es más fácil de interpretar y también se pueden comparar las tablas correspondientes en cuanto a los diferentes municipios. El argumento de que la tasa de valores perdidos es menor en un nivel general no justifica la práctica de omitir el ajuste, pues esa tasa puede diferir sustancialmente entre distintas áreas pequeñas.

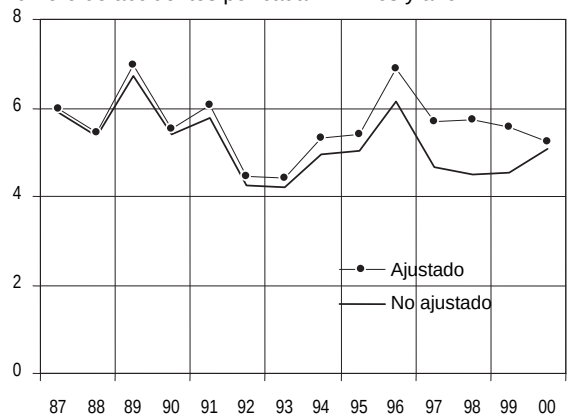
La no respuesta parcial varía con el tiempo

Se debe prestar atención a los valores perdidos dentro de las estadísticas basadas en registros y hacer los ajustes necesarios de una manera correspondiente a la realizada en las encuestas por muestreo. En el caso de las encuestas basadas en registros, es común que *no* haya ajustes por valores perdidos y simplemente se indique que hay un valor perdido. Si la magnitud de la tasa de valores perdidos varía con el tiempo y no se hacen correcciones, la comparabilidad a lo largo del tiempo será de baja calidad.

Ejemplo: Valores perdidos en el Registro de Pacientes

El Registro de Pacientes que mantiene el Consejo Nacional para la Salud y el Bienestar de Suecia contiene datos sobre las personas que han recibido atención hospitalaria. Se registra el diagnóstico, de modo que la no respuesta parcial para esta variable tiene como consecuencia una subestimación del número de pacientes con un diagnóstico específico. Las tasas de valores perdidos pueden variar drásticamente de un año a otro y de una región a otra, dependiendo de la eficiencia de los sistemas administrativos de los distintos hospitales. En el gráfico 8.4 vemos la diferencia entre una serie de tiempo *no* ajustada por no respuesta parcial en la variable *diagnóstico* y una serie de tiempo que sí se ajustó.

Gráfico 8.4 Accidentes por caída entre niños de 0 a 12 años en el condado de Norrbotten
Número de accidentes por cada mil niños y año

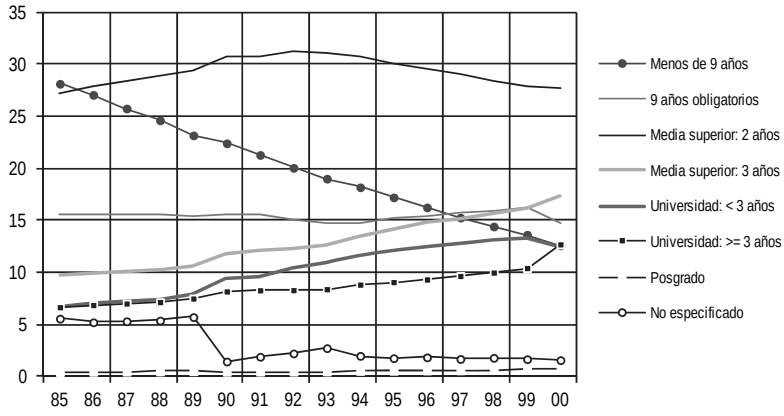


La trayectoria de la serie de tiempo se vuelve incorrecta cuando la tasa de valores perdidos varía con el tiempo.

Las comparaciones de valores incorrectos para las diferentes regiones son engañosas si las regiones presentan distintas tasas de valores perdidos.

Ejemplo: Valores perdidos en el Registro de Educación sueco
Las tasas de valores perdidos en el Registro de Educación han variado mucho con el tiempo y esas variaciones producen cambios aparentes en las distintas series de tiempo, como se observa en el gráfico 8.5. Para dar a los usuarios de estas estadísticas basadas en registros un panorama correcto de las trayectorias de las series de tiempo, se deben hacer ajustes por los efectos de los valores perdidos. Entre 1989 y 1990, las tasas de valores perdidos disminuyeron de 5.7 a 1.4%. Esto se debió a los datos que se recabaron para el Censo de Población de 1990. Salvo la educación obligatoria y el posgrado, todas las series aumentaron entre 1989 y 1990, pero se trata de incrementos en gran medida aparentes. También se observa un cambio de nivel en las series de tiempo entre 1999 y 2000 por los cambios en las clasificaciones educativas y la adición de nuevas fuentes. Analizaremos esto en la sección 8.3.

Gráfico 8.5 Efectos de los valores perdidos en una serie de tiempo del Registro de Educación. Población de 16 a 74 años de edad por nivel de escolaridad, 1985–2000
Porcentaje



8.1.3 Ajuste por valores perdidos con ponderadores

¿Qué métodos se pueden utilizar para hacer ajustes por valores perdidos? Mostraremos el método más simple, con el propósito de ilustrar el principio de *expansión directa* para datos ficticios del Registro de Educación. Podría hacerse un mejor ajuste para este registro si se

consideraran otras variables como edad, sexo, etc. Pero incluso este simple ajuste es mejor que ninguno.

Los ponderadores del registro se ajustan o calibran de acuerdo con la notación y los métodos presentados en la sección 7.5. La expansión directa significa que sólo se aplica una condición de calibración: el número total de observaciones debe corresponder al número de observaciones incluidas las que presentan valores perdidos.

En el siguiente gráfico se hace un ajuste para el Registro de Educación, donde el ponderador original $d_i = 1$ (0 para las observaciones con valor perdido).

Si la población consta de 6 386 015 personas, el número de valores perdidos es de 106 051 y hay datos para 6 279 964 personas, entonces el factor de ajuste g_i será 6 386 015/6 279 964 = 1.01689.

Gráfico 8.6 Ajuste por valores perdidos en el Registro de Educación 2001 con ponderadores

Persona	Sexo	Edad	Nivel de escolaridad	d_i	$d_i\, g_i = w_i$
NIP1	M	18	9 años obligatorios	1	1.01689
NIP2	F	72	Menos de 9 años	1	1.01689
NIP3	M	33	Media superior: 2 años	1	1.01689
NIP4	M	62	Media superior: 3 años	1	1.01689
NIP5	F	71	**Valor perdido**	0	0
NIP6	F	26	Universidad: ≥ 3 años	1	1.01689
NIP7	M	54	Posgrado	1	1.01689
NIP8	M	67	**Valor perdido**	0	0
NIP9	F	39	Menos de 9 años	1	1.01689
…	…	…	…	…	…
NIP6386015	M	53	Universidad: < 3 años	1	1.01689
Total:				**6 279 964**	**6 386 015**

Se usan los ponderadores w_i para calcular las estimaciones, que se ajustan para valores perdidos. A continuación se comparan las estimaciones no ajustadas y las ajustadas.

Gráfico 8.7 Tabla no ajustada

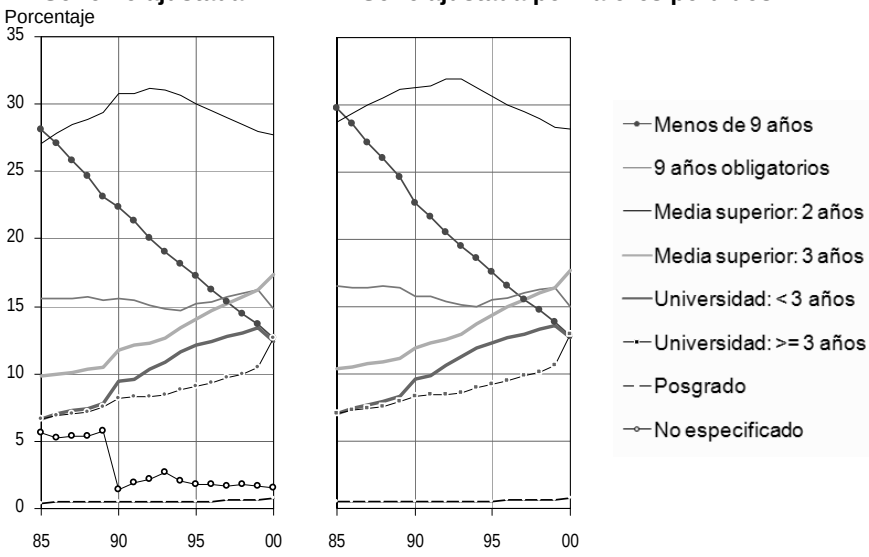
Registro de Educación 2001		
Nivel de escolaridad	Miles	% de la población
Menos de 9 años	755	11.8
9 años obligatorios	939	14.7
Media superior: 2 años	1 747	27.4
Media superior: 3 años	1 142	17.9
Universidad: < 3 años	802	12.6
Universidad: ≥ 3 años	848	13.3
Posgrado	48	0.7
No especificado	**106**	**1.7**
Población de 16 a 74 años	6 386	100.0

Gráfico 8.8 Tabla ajustada por valores perdidos

Registro de Educación 2001		
Nivel de escolaridad	Miles	% de la población
Menos de 9 años	767	12.0
9 años obligatorios	954	14.9
Media superior: 2 años	1 776	27.8
Media superior: 3 años	1 162	18.2
Universidad: < 3 años	816	12.8
Universidad: ≥ 3 años	862	13.5
Posgrado	48	0.8
Población de 16 a 74 años	6 386	100.0

Los ponderadores ajustados en una matriz de datos indican que todas las estimaciones se ajustaron de manera consistente y se mejoró la comparabilidad en el tiempo, lo que queda claro en el gráfico 8.9. Cuando la tasa de valores perdidos varía de un año a otro, los ponderadores ajustados diferirán según el año.

Gráfico 8.9 Población de 16 a 74 años por nivel de escolaridad, 1985–2000
A. Serie no ajustada B. Serie ajustada por valores perdidos



8.1.4 Ajuste por valores perdidos con imputación

Otra manera de hacer ajustes por no respuesta parcial consiste en formar *valores imputados* cuando hay valores de variables perdidos, es decir, los valores perdidos se sustituyen por valores sintéticos. Hay dos maneras distintas de formar estos valores:

– Aleatoriamente usando una o más distribuciones de probabilidad. Este método se aplica a las variables cualitativas.
– El valor se forma usando un modelo (determinístico) como se hace con las variables derivadas según se describe en la sección 6.2.3.

La ventaja de la imputación de valores de variables es que ya no es necesario calcular con ponderadores y las distribuciones del resto de las variables en el registro se mantienen sin cambios.

Valores imputados para variables cualitativas formadas aleatoriamente
En el gráfico 8.10 vemos cómo formar valores imputados para la variable *nivel de escolaridad*. Como se trata de un registro de personas, el valor para las personas NIP5 y NIP8 no debe imputarse de acuerdo con la legislación sueca, sino que se deben formar *observaciones sintéticas* que no tengan números de identificación personal.

La imputación, que corresponde a una expansión directa, se lleva a cabo de la siguiente manera: las observaciones con valores perdidos para *nivel de escolaridad* se usan para formar el mismo número de observaciones sintéticas. Con estas observaciones sintéticas se obtienen valores para la variable *nivel de escolaridad* de manera totalmente aleatoria. Dichos *niveles de escolaridad* elegidos al azar siguen la misma distribución que aquellos para los que se conocen los datos sobre el *nivel de escolaridad*. Se aumenta el registro con números aleatorios y luego se crea una matriz de datos sin números de identificación personal. Estos números aleatorios son valores para la variable técnica de uso interno.

Gráfico 8.10 Ajuste por valores perdidos en el Registro de Educación mediante imputación

A. Registro real, 2001

	Sexo	Edad	Nivel de escolaridad	Número aleatorio
NIP1	M	18	9 años obligatorios	0.7771
NIP2	F	72	Menos de 9 años	0.3168
NIP3	M	33	Media superior: 2 años	0.3096
NIP4	M	62	Media superior: 3 años	0.8667
NIP5	F	71	**Valor perdido**	**0.1749**
NIP6	F	26	Universidad: ≥ 3 años	0.4114
NIP7	M	54	Posgrado	0.1605
NIP8	M	67	**Valor perdido**	**0.5536**
NIP9	F	39	Menos de 9 años	0.5513
…	…	…	…	…
NIP6,386,015	M	53	Universidad: < 3 años	0.7828

B. Matriz de datos para análisis, 2001

Sexo	Edad	Nivel de escolaridad	Nivel de escolaridad imputado
M	18	9 años obligatorios	No
F	72	Menos de 9 años	No
M	33	Media superior: 2 años	No
M	62	Media superior: 3 años	No
F	71	**9 años obligatorios**	Sí
F	26	Universidad: ≥ 3 años	No
M	54	Posgrado	No
M	67	**Media superior: 3 años**	Sí
F	39	Menos de 9 años	No
…	…	…	…
M	53	Universidad: < 3 años	No

C. Distribución de probabilidad con base en la tabla de frecuencias del gráfico 8.8

Nivel de escolaridad	Proporción de la población	Proporción acumulada
Menos de 9 años	0.120	0.120
9 años obligatorios	0.149	0.269
Media superior: 2 años	0.278	0.547
Media superior: 3 años	0.182	0.729
Universidad: < 3 años	0.128	0.857
Universidad: ≥3 años	0.135	0.992
Posgrado	0.008	1.000
Población de 16 a 74 años	**1.000**	

Se hace una distribución uniforme de los números aleatorios del registro entre 0 y 1. A las personas con un número aleatorio entre 0 y 0.120 se les asigna el nivel *menos de 9 años*; a aquellas con un número aleatorio entre 0.120 y 0.269, el nivel *9 años obligatorios*, y así sucesivamente.

Las relaciones entre edad, sexo y nivel de escolaridad pueden servir para mejorar la imputación. Para las diferentes combinaciones de la categoría de edad y sexo, se aplican distintas distribuciones de frecuencias respecto al nivel de escolaridad. En el gráfico 8.11 se comparan estas tres distribuciones. Hay diferencias significativas entre ellas, lo que implica que es posible mejorar el ajuste por valores perdidos utilizando diferentes distribuciones cuando los valores se distribuyen aleatoriamente para las distintas combinaciones de sexo y edad.

Gráfico 8.11 Tabla de frecuencias por edad y sexo, Registro de Educación, 2001

Nivel de escolaridad	Participación acumulada Hombres de 65 a 74 años	Participación acumulada Mujeres de 65 a 74 años	Participación acumulada Ambos sexos de 16 a 74 años
Menos de 9 años	0.466	0.455	0.120
9 años obligatorios	0.507	0.532	0.269
Media superior: 2 años	0.700	0.821	0.547
Media superior: 3 años	0.837	0.858	0.729
Universidad: < 3 años	0.901	0.918	0.857
Universidad: ≥ 3 años	0.988	0.998	0.992
Posgrado	1.000	1.000	1.000

En el gráfico 8.12, se usa el mismo registro con los mismos números aleatorios anteriores. Sin embargo, en este caso los números aleatorios se convirtieron en nivel de escolaridad usando otras tablas de frecuencia. A las mujeres dentro de la categoría de edad de 65 a 74 años con un número aleatorio entre 0 y 0.455 se les asigna el nivel *menos de 9 años*, mientras que a los hombres en esa misma categoría de edad con un número aleatorio entre 0.507 y 0.700, se les asigna el nivel *media superior: 2 años*. De la misma manera, a las personas más jóvenes con *valor perdido* se les asigna un valor imputado mediante tablas de frecuencia para su categoría de edad y sexo.

Gráfico 8.12 Ajuste para valores perdidos en el Registro de Educación con imputación

A. Registro real, 2001

	Sexo	Edad	Nivel de escolaridad	Número aleatorio
NIP1	M	18	9 años obligatorios	0.7771
NIP2	F	72	Menos de 9 años	0.3168
NIP3	M	33	Media superior: 2 años	0.3096
NIP4	M	62	Media superior: 3 años	0.8667
NIP5	F	71	**Valor perdido**	**0.1749**
NIP6	F	26	Universidad: ≥ 3 años	0.4114
NIP7	M	54	Posgrado	0.1605
NIP8	M	67	**Valor perdido**	**0.5536**
NIP9	F	39	Menos de 9 años	0.5513
...	...	...	...	...
NIP6386015	M	53	Universidad: < 3 años	0.7828

B. Matriz de datos para análisis, 2001

Nivel de escolaridad imputado	Sexo	Edad	Nivel de escolaridad
No	M	18	9 años obligatorios
No	F	72	Menos de 9 años
No	M	33	Media superior: 2 años
No	M	62	Media superior: 3 años
Sí	F	71	**Menos de 9 años**
No	F	26	Universidad: ≥ 3 años
No	M	54	Posgrado
Sí	M	67	**Media superior: 2 años**
No	F	39	Menos de 9 años
...	...	...	...
No	M	53	Universidad: < 3 años

El *NIP5*, una mujer de 71 años, tiene como nivel de escolaridad *menos de 9 años*, que difiere de la imputación del gráfico 8.10, donde se le asigna *9 años obligatorios.* El nivel imputado para el *NIP8* también cambió a un periodo de educación más breve.

¿Cuándo es apropiado usar valores imputados aleatoriamente?

El método anterior es apropiado cuando se describe una variable cualitativa con no respuesta parcial, posiblemente dividida en diferentes categorías, como edad, sexo y región. Después de una imputación de alta calidad, los niveles son más comparables entre categorías y a lo largo del tiempo si no se hacen ajustes por valores perdidos.

Cuando se estudia la relación entre la variable *y* y una variable cualitativa *x*, donde la variable *x* presenta una no respuesta parcial, no se deben usar valores imputados aleatoriamente para la variable *x*. Por ejemplo, no sería apropiado usar niveles de escolaridad imputados aleatoriamente al describir el salario mensual promedio para los distintos niveles de escolaridad. En ese caso, sería mejor calcular el salario promedio sólo para las personas sobre las que se conoce el nivel de escolaridad.

Valores imputados usando un modelo determinístico

En el método de imputación aplicado en el gráfico 8.12, se usa la relación entre las variables *x*, edad y sexo, y la variable *y*, nivel de escolaridad. Sin embargo, los valores imputados también se forman *aleatoriamente*. Para una combinación particular de edad y sexo, el nivel de escolaridad no está determinado con exactitud, sino de manera aleatoria. Enseguida presentamos algunos ejemplos de métodos de imputación donde el valor de las variables *x* determina *exactamente* los valores imputados. Los modelos usados para este tipo de imputación reciben el nombre de *modelos determinísticos*.

En la sección 6.2.3 vimos cómo formar variables derivadas con modelos causales determinísticos. Los valores de variables imputados se pueden formar de manera similar. La diferencia es que los valores de variables derivados se calculan para *todos* los objetos de la matriz de datos, mientras que los valores de variables imputados sólo se forman para aquellos objetos con valores perdidos a causa de la no respuesta parcial.

Durante el trabajo de validación se identifica si hay valores perdidos o ilógicos, los cuales deben rechazarse. Esto hace que el cálculo de valores imputados guarde estrecha relación con el trabajo de validación. Los estudios de caso sobre validación de la sección 6.3 contienen varios ejemplos de métodos de imputación.

Al validar el Registro de Ingresos y Tributación, se descubre que no se ha notificado la asistencia social para algunos municipios (véase sección 6.3.1). Por lo tanto, se imputaron los valores del año anterior a los hogares de estos municipios. Para la imputación se utiliza

un modelo simple: *asistencia del año en curso = asistencia del año anterior*. De modo que en el nivel de los hogares pueden ocurrir *errores de modelación* o *de imputación* si la asistencia de ese año es distinta a la del año anterior. Se debe intentar usar modelos que produzcan los menores errores de imputación posibles.

Cuando los requisitos de calidad son tan estrictos que los errores de imputación resultan inaceptables, se recomienda llevar a cabo una *captación especial de datos*. Entonces los objetos que carecen de valores para una variable importante pueden dar los valores perdidos por medio de un cuestionario o una entrevista. En la sección 6.3.2 describimos la validación de los datos anuales para estados de resultados. Para todos los datos sobre estados de resultados, se debe indicar una identidad de establecimiento. Cuando son datos perdidos o se consideran ilógicos, se debe entrar en contacto con el empleador.

La validación de las declaraciones de ingresos nos da ejemplos de los diferentes tipos de métodos de imputación:

– Los datos sobre el número de empleados durante todo el año se toman de los informes anuales. Si estos datos están perdidos, se forman valores imputados mediante una estimación del número de empleados durante todo el año dividiendo la suma salarial de la empresa entre el promedio por empleado durante todo el año en la rama industrial. El salario promedio para la rama industrial se calculó usando aquellas empresas sobre las que se conocen el número de empleados durante todo el año y las sumas salariales.
– La población de registro en las Estadísticas Empresariales Estructurales carece de valores para variables económicas en el caso de algunas empresas. Los datos sobre la rama industrial y el número de empleados durante todo el año se importaron del Registro Empresarial. En el caso de empresas sobre las que se conocen la rama industrial, el número de empleados durante todo el año y las variables económicas, se elaboran tablas con valores medios para las diferentes variables económicas, por rama industrial y número de empleados durante todo el año. Para determinados valores de la rama industrial y el número de empleados durante todo el año, estas tablas son una forma de modelo que indica la manera en que se deben generar los valores imputados a partir de los valores promedio calculados.

8.1.5 Valores perdidos en un sistema de registros

Cuando se integran diferentes registros y se importan variables de un registro a otros, también se importan fallas en la calidad, como los valores perdidos.

Por ejemplo, en el Registro Empresarial se crea la variable *rama industrial* y luego se importa a otros registros empresariales, registros de actividades, registros sobre personas y también registros inmobiliarios. Esto significa que no basta con hacer ajustes por los valores perdidos en la variable *rama industrial* en un solo registro; se debe aplicar el método seleccionado y hacer los ajustes para esa variable en todo el sistema de registros de una manera consistente.

Después de tratar de reducir la tasa de no respuesta parcial usando más fuentes, y quizá también captando información de ciertas categorías de objetos, las estimaciones de las estadísticas basadas en registros se deben ajustar por valores perdidos. Aquí comparamos los dos métodos de ajuste: con ponderadores o con imputación de valores.

Ajuste por valores perdidos con ponderadores en un sistema de registros

Los registros de Población, Educación y Empleo se relacionan con la población al 31 de diciembre de un año particular. No hay no respuesta parcial en el Registro de Población, mientras que el Registro de Educación y el Registro de Empleo sí contienen no respuesta

parcial, el primero en la variable nivel de escolaridad y el segundo tanto en el nivel de escolaridad como en la clasificación industrial.

Si cada registro se ajusta por separado con ponderadores para los valores perdidos, los ponderadores para la misma persona serán distintos para cada uno de los registros. Esto se ilustra en el gráfico 8.13. Las estadísticas de los tres registros serán entonces inconsistentes, por ejemplo, el número de hombres de 18 años será distinto (el NIP1 tiene diferentes ponderadores en las tablas A, B y C del gráfico 8.13).

Para que las estadísticas de los diferentes registros relacionados con la misma población sean consistentes, los ponderadores se deben calcular conjuntamente y se deben usar los mismos ponderadores para todos los registros. Esto puede ser difícil de lograr. Nuestra conclusión es que los ajustes por valores perdidos con ponderadores ocasionarán problemas para la coordinación y consistencia dentro del sistema de registros.

Gráfico 8.13 Ajuste por valores perdidos utilizando ponderadores en un sistema de registros

A. Reg. de Población

Persona	Sexo	Edad	d_i
NIP1	M	18	1
NIP2	F	72	1
NIP3	M	33	1
NIP4	M	62	1
NIP5	F	71	1
NIP6	F	26	1
NIP7	M	54	1
NIP8	M	67	1
NIP9	F	39	1
...	...	...	...

B. Registro de Educación

NIP	Nivel de escolaridad	$d_i g_i$
NIP1	9 años obligatorios	1.01689
NIP2	Menos de 9 años	1.01689
NIP3	Media superior: 2 años	1.01689
NIP4	Media superior: 3 años	1.01689
NIP5	**Valor perdido**	**0**
NIP6	Universidad: ≥ 3 años	1.01689
NIP7	Posgrado	1.01689
NIP8	**Valor perdido**	**0**
NIP9	Menos de 9 años	1.01689
...	...	...

C. Registro de Empleo (16 a 64 años)

NIP	Rama industrial	Nivel de escolaridad	$d_i g_i$
NIP1	DM[1]	9 años obligatorios	1.02930
-	-	-	-
NIP3	**Perdido**	Media superior: 2 años	**0**
NIP4	DK	Media superior: 3 años	1.02183
-	**-**	**-**	**-**
NIP6	DB	Universidad: ≥ 3 años	1.02326
NIP7	DK	Posgrado	1.02326
-	-	**-**	**-**
NIP9	DM	Menos de 9 años	1.02930
...	...	...	...

Nota: Hay tres personas, NIP2, NIP5 y NIP8, que no tienen un trabajo remunerado de acuerdo con el Registro de Empleo y su edad tampoco entra en el intervalo 16 a 64 años. Los ponderadores $d_i g_i$ del gráfico 8.13 B son los mismos del gráfico 8.6 y los ponderadores $d_i g_i$ en C se calcularon comparando el número de personas en diferentes celdas de los gráficos 8.1 y 8.3. Por ejemplo, 281/273=1.02930.

Ajuste por valores perdidos con imputación en un sistema de registros

Si diferentes registros del sistema se ajustan por valores perdidos usando la imputación como se describe en la sección 8.1.4, las estadísticas de los diferentes registros podrían ser completamente consistentes. Al mismo tiempo que se importa una variable, se importan los números aleatorios (o valores imputados) que se usan en el registro original. Entonces se pueden hacer imputaciones consistentes entre diferentes registros.

En el ejemplo siguiente vemos que es posible importar al Registro de Empleo el nivel de escolaridad del Registro de Educación y la clasificación industrial del Registro Empresarial. A continuación los valores perdidos de estos registros se pueden remplazar por los valores imputados de una manera consistente.

[1] Fabricación de : DM = Vehículos de motor; DK = Maquinaria; DB = Textiles; DA = Productos alimenticios; DH = Industria de la transformación del caucho y materias plásticas; DJ = Metalurgia, fabricación de productos metálicos; DK = Industria de la construcción de maquinaria y equipo mecánico.

Gráfico 8.14 Ajustes por valores perdidos en el Registro de Educación por medio de imputación

A. Registro real

Persona	Sexo	Edad	Nivel de escolaridad	Número aleatorio
NIP1	M	18	9 años obligatorios	0.7771
NIP2	F	72	Menos de 9 años	0.3168
NIP3	M	33	Media superior: 3 años	0.3096
NIP4	M	62	Media superior: 3 años	0.8667
NIP5	F	71	**Valor perdido**	**0.1749**
NIP6	F	26	Universidad: ≥ 3 años	0.4114
NIP7	M	54	Posgrado	0.1605
NIP8	M	67	**Valor perdido**	**0.5536**
…	…	…	…	…

B. Matriz de datos para análisis

Sexo	Edad	Nivel de escolaridad	Nivel de escolaridad imputado
M	18	9 años obligatorios	No
F	72	Menos de 9 años	No
M	33	Media superior: 2 años	No
M	62	Media superior: 3 años	No
F	71	**9 años obligatorios**	Sí
F	26	Universidad: ≥ 3 años	No
M	54	Posgrado	No
M	67	**Media superior: 3 años**	Sí
…	…	…	…

Gráfico 8.15 Ajuste por valores perdidos en el Registro Empresarial por medio de imputación

A. Registro real

Empresa	Rama industrial	Número aleatorio
UJ1	DB	0.0316
UJ2	DK	0.6444
UJ3	**Valor perdido**	**0.3978**
UJ4	DA	0.2846
UJ5	DK	0.2044
…	…	…

B. Matriz de datos para análisis

Rama industrial	Rama industrial imputada
DB	No
DK	No
DM	Sí
DA	No
DK	No
…	…

Gráfico 8.16 Ajuste para valores perdidos en el Registro de Empleo con imputación

A. Registro real

Persona	Empresa	Rama industrial	Núm. aleatorio para rama industrial	Nivel de escolaridad	Núm. aleatorio para nivel de escolaridad
NIP1	UJ5	DK	0.2044	9 años obligatorios	0.7771
NIP2	-	-	-	Menos de 9 años	0.3168
NIP3	UJ3	**Perdido**	**0.3978**	Media superior: 2 años	0.3096
NIP4	UJ2	DK	0.6444	Media superior: 3 años	0.8667
NIP5	-	**-**	**-**	**Valor perdido**	**0.1749**
NIP6	UJ1	DB	0.0316	Universidad: ≥ 3 años	0.4114
NIP7	UJ5	DK	0.2044	Posgrado	0.1605
NIP8	-	-	-	**Valor perdido**	**0.5536**
…	…	…	…	…	…

B. Matriz de datos para análisis

Rama industrial	Rama industrial imputada	Nivel de escolaridad	Nivel de escolaridad imputado
DK	No	9 años obligatorios	No
-	-	Menos de 9 años	No
DM	Sí	Media superior: 2 años	No
DK	No	Media superior: 3 años	No
-	-	**9 años obligatorios**	Sí
DB	No	Universidad: ≥ 3 años	No
DK	No	Posgrado	No
-	-	**Media superior: 3 años**	Sí
…	…	…	…

Conclusiones

Las conclusiones a las que llegamos luego de esta exposición son que se debe hacer el ajuste para valores perdidos, los ajustes deben coordinarse y la imputación es el método más apropiado para el ajuste por no respuesta parcial en un sistema de registros.

Dentro del sistema, los registros de Educación y Empresarial serán responsables de los ajustes por no respuesta parcial de *educación* y *rama industrial*, respectivamente. Después otros registros deben aplicar estos ajustes.

8.2 MÉTODOS DE ESTIMACIÓN PARA CORREGIR LA SOBRECOBERTURA

En esta sección explicamos cómo usar la calibración para corregir la sobrecobertura en un registro. Hoy en día esto no suele corregirse, pero podrían usarse los métodos que presentamos aquí para corregir estas fuentes de error.

Sobrecobertura en el Registro de Población

El primer indicio de que hay sobrecobertura en el Registro de Población de Suecia vino de los estudios demográficos sobre mortalidad. En algunas categorías de personas nacidas en el extranjero, la mortalidad era anormalmente baja. Además, se observó que la proporción de familias, sobre las que no había información sobre ingreso disponible, era elevada entre ciertas categorías de inmigrantes.

Greijer (1995, 1996, 1997a, 1997b), quien analizó la no respuesta en las Encuestas sobre Fuerza Laboral y en un censo sobre personas nacidas en el extranjero basado en un cuestionario postal, estimó la sobrecobertura en el Registro de Población sueco. A partir de esta información, se estimó la sobrecobertura en diferentes categorías de personas nacidas en el extranjero.

Los datos de los registros estadísticos también pueden servir como indicios de sobrecobertura. Por ejemplo, una persona nacida en el extranjero sin ingresos en ningún registro puede haberse ido de Suecia sin notificarlo a las autoridades hacendarias.

La sobrecobertura puede ocasionar errores graves en las estadísticas basadas en registros. Por ejemplo, puede ser engañoso el ingreso promedio de quienes nacieron en diferentes países. En el caso de personas nacidas en ciertos países, la subestimación puede rondar el 20 por ciento.

¿Cómo debemos controlar la sobrecobertura y mejorar la calidad? La estrategia para corregir errores derivados de la sobrecobertura puede incluir las siguientes medidas:

1. Siendo cuidadosos durante la macro validación, es posible detectar estimaciones ilógicas en las estadísticas basadas en registros. Uno se debe preguntar si la sobrecobertura podría ser la causa de estas estimaciones extremas.
2. Si se sospecha que hay una sobrecobertura, se puede recurrir a las encuestas por muestreo disponibles y otras fuentes como apoyo para estimarla.
3. Se puede estimar la sobrecobertura para diferentes categorías en el registro una vez que se ha captado información suficiente sobre la magnitud y la naturaleza de dicha sobrecobertura.
4. A continuación, se pueden ajustar los ponderadores para corregir la sobrecobertura estimada. Antes del ajuste todos los ponderadores son iguales a 1; después del ajuste, los ponderadores para las diferentes categorías en las que hay sobrecobertura serán menores a 1. Cuando muchas variables describen una sobrecobertura, se deben usar los *métodos de calibración* (sección 7.5) para ajustar los ponderadores.
5. Los ponderadores ajustados se guardan en el registro base (en este caso, el Registro de Población).
6. Se aplicarán los ponderadores en el resto de los productos estadísticos que usan el registro base. De esta forma, en todas las estadísticas producidas se corregirá de manera consistente el efecto estimado de sobrecobertura.

Ejemplo: En un registro (ficticio) integrado por mil personas nacidas en el extranjero, la sobrecobertura alcanza 10%, es decir, los mil objetos del registro corresponden a 900 personas en realidad. Además, suponemos que se ha estimado que hay sobrecobertura en

diferentes categorías de acuerdo con el gráfico 8.17 y que usamos esta información para calcular los ponderadores ajustados.

Gráfico 8.17 Sobrecobertura estimada para diferentes categorías

		Número de personas antes de la corrección (1)	Sobrecobertura estimada (2)	Núm. de personas después de la corrección por sobrecobertura (3)
País de nacimiento	Europa	*584*	6.7%	**545**
	No Europa	*416*	14.7%	**355**
	Total	1 000	10.0%	900
Años en Suecia	Pocos	*819*	7.2%	**760**
	Muchos	181	22.7%	140
	Total	1 000	10.0%	900
Ingresos	Bajos	101	40.6%	60
	Altos	*899*	6.6%	**840**
	Total	1 000	10.0%	900

En el gráfico 8.17, parece que tenemos seis expresiones para el número de personas, pero en realidad sólo hay cuatro expresiones porque a partir de esos cuatro números, en negritas, podemos calcular el resto. Estos cuatro números después de la corrección y los números correspondientes antes de la corrección (en cursivas) se usarán enseguida como condiciones de calibración.

Calibración de ponderadores

El gráfico 8.18 muestra como se calculan los ponderadores ajustados usando la metodología de calibración descrita en la sección 7.5. En las primeras cinco columnas aparece el registro original, mientras que las columnas $x_{1i} - x_{4i}$ contienen la información que se usará para la calibración. En los cálculos, se usan los vectores $\boldsymbol{x}_i'$, uno por renglón. Para $i=1$, que corresponde a *NIP1*, $\boldsymbol{x}_1' = (1 \quad 0 \quad 1 \quad 1)$. Las cuatro condiciones de calibración definen el vector $\boldsymbol{t}_x$ con las frecuencias absolutas correctas y el vector $\boldsymbol{\Sigma} d_i \boldsymbol{x}_i$ con las frecuencias absolutas incorrectas. Estas cuatro condiciones se encuentran en el gráfico 8.17.

Gráfico 8.18 Registro con ponderadores calibrados

NIP	País de nacimiento	Años en Suecia	Ingresos	Ponderadores d_k	x_{1i} País Europa	x_{2i} País No Europa	x_{3i} Pocos años	x_{4i} Altos ingresos	Ponderador ajustado, w_i
1	Europa	Pocos	Altos	1	1	0	1	1	0.992
2	No Europa	Pocos	Altos	1	0	1	1	1	0.916
3	Europa	Pocos	Bajos	1	1	0	1	0	0.657
4	Europa	Pocos	Altos	1	1	0	1	1	0.992
5	No Europa	Muchos	Altos	1	0	1	0	1	0,770
6	No Europa	Pocos	Bajos	1	0	1	1	0	0.581
...	...	...	...	...	...	...	...	...	...
1,000	No Europa	Pocos	Altos	1	0	1	1	1	0.916
Total				**1,000**	**584**	**416**	**819**	**899**	**900**

Condiciones

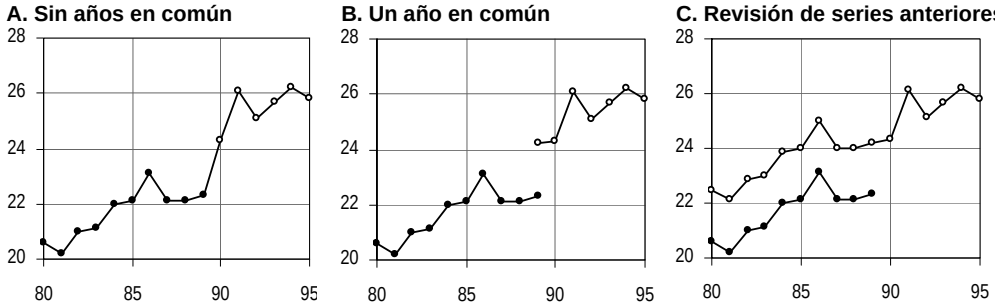
t_x	$\sum d_i \boldsymbol{x}_i$
545	*584*
355	*416*
760	*819*
840	*899*

La sobrecobertura en el Registro de Población sueco podría parecer de importancia menor: de nueve millones de registros, se sospecha que sólo 50 000 son sobrecobertura, es decir, 0.6%. Sin embargo, este error dificulta muchas aplicaciones importantes del registro, por ejemplo, para comparar las diferentes clases de inmigrantes. Se desconoce la identidad de estas 50 000 personas, pero se sabe que pertenecen a ciertas categorías de personas. Con el uso de ponderadores para estas categorías en el Registro de Población, se pueden reducir los errores de sobrecobertura.

8.3 MÉTODOS PARA CORREGIR CAMBIOS DE NIVEL EN LAS SERIES DE TIEMPO

Los registros estadísticos de distintos años se usan para producir series de tiempo. Se pueden producir miles de series de tiempo mediante un registro estadístico; pero cuando cambia el *sistema administrativo*, que es la fuente de las modificaciones de ese registro, puede haber cambios de nivel en la series de tiempo. Los cambios efectuados por el productor de las estadísticas también pueden originar cambios de nivel. Los *sistemas de clasificación*, como los de clasificación industrial, ocupación o nivel de escolaridad, se revisan y actualizan de manera periódica, estos cambios también traen consigo cambios de nivel en las series de tiempo. Los *cambios de calidad* a causa de la revisión de las metodologías o del uso de nuevas fuentes de información también pueden dar lugar a cambios de nivel.

Gráfico 8.19 Tres maneras de presentar cambios de nivel en series de tiempo



El gráfico 8.19A muestra una situación en la que el productor no ha hecho esfuerzo alguno para medir el efecto de un cambio de nivel; en el gráfico B, el productor ha llevado a cabo cálculos dobles para 1989 a fin de ilustrar el cambio de nivel. En el gráfico C, también se recalculó la serie de tiempo para años previos.

No se debe pasar por alto el cambio de nivel, como se muestra en la gráfica de la izquierda, pues entonces los usuarios tendrán que interpretar los datos completamente solos. No basta con mencionar que hay un cambio de nivel, hay que explicar cómo se deben interpretar los datos. En las publicaciones, a veces se utiliza una observación de tipo *"Los resultados de años anteriores se deben interpretar con cautela"*. No obstante, ningún usuario se sentirá satisfecho con esto porque no queda claro qué significa interpretar los datos "con cautela".

Como requisito mínimo, el productor responsable debe llevar a cabo cálculos como los del gráfico 8.19B, es decir, mostrar el efecto del año en el que ocurrió el cambio. Pero la mejor solución aparece en el gráfico 8.19C, donde las series de tiempo se vinculan para producir series más largas en las que se minimiza el efecto del cambio de nivel.

Hacer correcciones para los cambios de nivel en una serie de tiempo se conoce como *vinculación de series de tiempo*. Diferenciamos entre *vinculación en el nivel macro* y *vinculación en el nivel micro*. Es común hacer la vinculación en el nivel macro, es decir, trabajar exclusivamente con los valores agregados de las series de tiempo y tratar de corregirlos los efectos de los cambios de nivel de las series de tiempo. Este método tiene la desventaja de que puede ser difícil asegurar la consistencia entre varias series vinculadas. Por ejemplo, la suma de las series revisadas para hombres y mujeres debe coincidir con la serie revisada de toda la población. Cuando hay muchas series, la vinculación en el nivel macro se vuelve problemática ya que se deben satisfacer muchos requisitos de consistencia.

Por lo general, los responsables de un registro estadístico pueden vincular *un número limitado* de las series de tiempo más importantes cuando ha ocurrido un cambio que afecta el registro. Pero ¿qué hacer con las otras miles de series que posiblemente también resulten afectadas? Supongamos que se han estimado cambios de nivel para cien series, pero el registro estadístico en realidad genera un total de 10,000 series, ¿qué se puede hacer con las otras 9,900 series?

La vinculación en el nivel micro no presenta estas desventajas, pues todas las series de tiempo posibles se vinculan al mismo tiempo y serán consistentes. Este método se lleva a cabo calculando los ponderadores corregidos en las matrices de datos para los años o periodos que se revisarán. Los ponderadores originales indican los valores de las series de tiempo anteriores, mientras que los ponderadores revisados se pueden usar para calcular las series vinculadas.

A continuación veremos un procesamiento de registros que permite calcular los valores revisados de las series de tiempo para periodos anteriores. Para usar ese método, primero se estimaron los efectos del correspondiente al periodo anterior haciendo una vinculación en el nivel macro para un número limitado de series.

8.3.1 Estimación de los efectos de los cambios de nivel: vinculación en el nivel macro

Si se hacen modificaciones dentro del sistema administrativo, el productor de las estadísticas debe dar seguimiento a las modificaciones efectuadas y medir los cambios de nivel. Esto se puede hacer con información de los registros administrativos o usando las encuestas por muestreo. Entonces se pueden estimar los cambios de nivel para las series de tiempo más importantes.

En el año 2000, además de que se introdujo un nuevo sistema de clasificación educativa en el Registro de Educación sueco, se usaron nuevas fuentes por primera vez. Se analizaron los efectos del cambio 1999/2000 en el registro. Se comparó la población entre 16 y 74 años por tipo de educación: según la antigua clasificación de las fuentes anteriores y según la nueva clasificación cuando se habían agregado las fuentes nuevas. Añadimos tres columnas en el gráfico 8.20, donde también hicimos ajustes por valores perdidos.

Gráfico 8.20 Estimación de los cambios de nivel en 1999 para los diferentes tipos de educación

Personas de 16 a 74 años por nivel de escolaridad	Sin ajuste por valores perdidos:		Después del ajuste por valores perdidos, por medio de expansión directa:		
Nivel de escolaridad	**1999 anteriores**	**1999 nuevos**	**1999 anteriores**	**1999 nuevos**	**Diferencia**
Menos de 9 años	863 332	848 289	877 948	862 428	–15 520
9 años obligatorios	1 025 322	926 996	1 042 680	942 446	–100 234
Media superior: 2 años	1 768 378	1 798 085	1 798 316	1 828 054	29 738
Media superior: 3 años	1 024 001	1 091 800	1 041 337	1 109 997	68 660
Universidad: < 3 años	846 876	748 071	861 213	760 539	–100 674
Universidad: ≥ 3 años	660 142	776 418	671 318	789 359	118 041
Posgrado	41 619	41 619	42 324	42 313	–11
No especificado	105 466	103 858	0	0	0
Población de 16 a 74 años	**6 335 136**	**6 335 136**	**6 335 136**	**6 335 136**	**0**

¿Cómo se puede usar ahora esta información para corregir los valores anteriores de las series de tiempo? Un método común es el *método del cociente*, que en este caso consiste en multiplicar todos los valores anteriores de las series de tiempo por un cociente entre los valores nuevos y anteriores para 1999. Para la serie *Personas con menos de 9 años de*

escolaridad, el cociente es 862 428/877 948 = 0.982. De la misma manera, todos los valores anteriores para las otras seis series de tiempo podrían multiplicarse por los cocientes que les correspondieran. Sin embargo, este método para vincular series no ofrece suficiente consistencia. Las siete series de tiempo no suman la población total para los años anteriores a 1999.

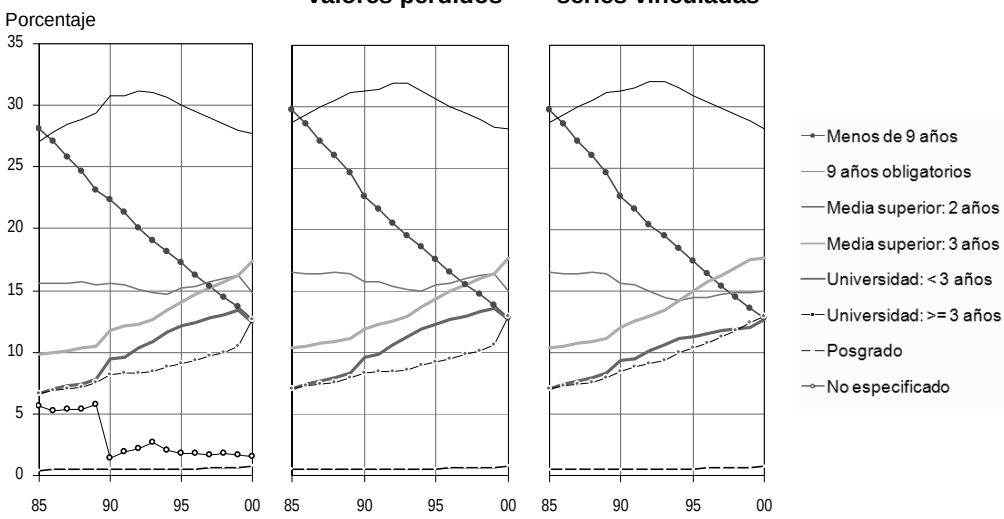
Si las series se van a sumar, se debe usar el modelo aditivo en vez del multiplicativo como método para obtener el cociente. En el gráfico 8.21 se muestra un posible método para vincular las series usando un modelo aditivo. Para 1999, las series se corrigen de acuerdo con las diferencias aplicables a ese año (gráfico 8.20). Entonces consideramos que estas diferencias han ocurrido paulatinamente durante el periodo 1990–1999, de modo que reducimos las diferencias de nivel un décimo cada año. Estas correcciones son consistentes, pues no cambia el número total de personas entre 16 y 74 años. No están corregidos los años anteriores a 1990 y posteriores a 1999. Ilustramos los principios del método sólo con estas siete series de tiempo vinculadas. Se debe hacer una vinculación aún mejor en la que también se preste atención a los campos *nivel de escolaridad*, *edad*, *sexo*, etcétera, de modo que alrededor de 100 series se vinculen primero en el nivel macro. Las series vinculadas se usarán como condiciones de calibración en la sección 8.3.2.

Gráfico 8.21 Correcciones en las series de tiempo aplicando el modelo aditivo, número de personas

	Menos de 9 años	9 años obligatorios	Media superior: 2 años	Media superior: 3 años	Universidad: < 3 años	Universidad: ≥ 3 años	Posgrado	Total
1990	-1 552	-10 023	2 974	6 866	-10 067	11 804	-1	0
1991	-3 104	-20 047	5 948	13 732	-20 135	23 608	-2	0
1992	-4 656	-30 070	8 921	20 598	-30 202	35 412	-3	0
1993	-6 208	-40 094	11 895	27 464	-40 270	47 216	-4	0
1994	-7 760	-50 117	14 869	34 330	-50 337	59 020	-5	0
1995	-9 312	-60 140	17 843	41 196	-60 404	70 824	-7	0
1996	-10 864	-70 164	20 817	48 062	-70 472	82 629	-8	0
1997	-12 416	-80 187	23 790	54 928	-80 539	94 433	-9	0
1998	-13 968	-90 210	26 764	61 794	-90 607	106 237	-10	0
1999	**-15 520**	**-100 234**	**29 738**	**68 660**	**-100 674**	**118 041**	**-11**	**0**

En el gráfico 8.22C vemos el efecto de esta vinculación de series de tiempo. Han desaparecido los mayores cambios de nivel en el tiempo para educación obligatoria, 3 años de media superior y ambas series universitarias. Al comparar las tablas A, B y C del gráfico 8.22, se observa el efecto de las diferentes etapas en el mejoramiento de la comparabilidad con el paso de los años.

Gráfico 8.22 Población de 16 a 74 años por nivel de escolaridad, 1985–2000



8.3.2 Vinculación de series de tiempo en el nivel micro con ponderadores

En esta situación también es posible usar métodos basados en la calibración de ponderadores. La metodología de calibración se aplica en la sección 7.5, donde mostramos cómo corregir la no respuesta parcial. Se puede seguir un procedimiento similar para corregir los cambios de nivel en las series de tiempo. Continuando con el ejemplo del Registro de Educación, el proceso se puede dividir en las siguientes etapas:

– Supongamos que 1999 es el primer año después de los cambios y que existen cien series de tiempo importantes de acuerdo con los niveles tanto anteriores como nuevos para 1999 y años previos. Por consiguiente, se han vinculado cien series para el periodo 1985–1999, de la misma manera en que vinculamos las siete series anteriores. Los cien valores vinculados para cada año son completamente consistentes entre sí, lo que significa que todos los totales concuerdan, por ejemplo, el número de hombres + el número de mujeres = número total de personas con una característica particular.

– Entonces es posible calcular el ponderador ajustado w_i de los objetos del registro para cada año, de modo que puedan calcularse nuevos valores de las series de tiempo comparables con los nuevos valores de las series de tiempo para 1999 y años posteriores. Los cien valores recalculados para un año particular se usan como vector $\boldsymbol{t}_x$ en el proceso de calibración. Para *cada año*, se calculan los ponderadores ajustados w_i descritos en la sección 7.5.

– Con los ponderadores w_i, se pueden estimar de manera consistente todos los valores posibles de la series de tiempo para cada año y las estimaciones serán comparables con los valores relativos a 1999 y años posteriores.

Hemos usado este método en la Oficina de Estadística de Suecia. Todas las encuestas de fuerza laboral del periodo 1987–1992 se calibraron a fin de hacer ajustes por los cambios introducidos en 1993. Wallgren (1998) describe con mayor detalle este procedimiento. Se usaron poco más de cien series vinculadas como condiciones de calibración y, luego de una verificación completa de la forma en que el método genera otras series vinculadas, se

descubrió que, con las ecuaciones de calibración apropiadas, las otras series también podían vincularse de manera aceptable.

A continuación ilustramos cómo funciona esta calibración continuando con el ejemplo del Registro de Educación, pero usando sólo las siete series vinculadas de la sección anterior. Esto establece siete condiciones de calibración. La tabla siguiente está tomada del gráfico 8.20, en el que se aborda la vinculación en el nivel macro.

Gráfico 8.23 Cambios de nivel estimados

Nivel de escolaridad	Nuevos para 1999: t_x	Anteriores para 1999: $\sum d_i\, g_i\, x_i$	Cambio de nivel: $t_x - \sum d_i\, g_i\, x_i$
Menos de 9 años	862 428	877 948	-15 520
9 años obligatorios	942 446	1 042 680	-100 234
Media superior: 2 años	1 828 054	1 798 316	29 738
Media superior: 3 años	1 109 997	1 041 337	68 660
Universidad: < 3 años	760 539	861 213	-100 674
Universidad: ≥ 3 años	789 359	671 318	118 041
Posgrado	42 313	42 324	-11

En el gráfico 8.24 vemos cómo se calculan los ponderadores ajustados usando la metodología de calibración. Las columnas $x_{1i} - x_{7i}$ muestran los vectores $\boldsymbol{x}_i'$, un vector por renglón. Para *i=1*, como en el caso de *NIP1*, $\boldsymbol{x}_1' =$ (1 0 0 0 0 0 0).

Los ponderadores originales $d_i = 1$ se calibraron primero a fin de hacer ajustes por no respuesta parcial. Luego de esta primera calibración, obtenemos los ponderadores $d_i\, g_i$, que se calibran de nuevo con objeto de hacer ajustes por cambios de nivel. La última columna del registro muestra el ponderador ajustado w_k, que se calcula de la siguiente forma:

1. Se calculan $\boldsymbol{T} = \sum d_i\, g_i\, \boldsymbol{x}_i\, \boldsymbol{x}'_i$ y $\boldsymbol{T}^{-1}$, durante la calibración se deben corregir los ponderadores $d_i\, g_i$, donde i = 1, 2, 3,…, 6335136.
   $\boldsymbol{T}$ es la matriz con los totales al cuadrado y de los productos, en este caso una matriz de 7×7. Como $x_{1i} - x_{7i}$ son 1 o 0, se trata de una matriz que tiene en la diagonal principal los números no corregidos para cada nivel de escolaridad y ceros fuera de la diagonal principal:

$$\boldsymbol{T} = \begin{bmatrix} 877\,948 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1\,042\,680 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1\,798\,316 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1\,041\,337 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 861\,213 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 671\,318 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 42\,324 \end{bmatrix}$$

2. Se calcula el vector $\boldsymbol{\lambda}$: $\boldsymbol{\lambda} = \boldsymbol{T}^{-1}(\boldsymbol{t}_x - \sum d_i\, g_i\, \boldsymbol{x}_i)$
   El vector $\boldsymbol{t}_x$ constituye las siete expresiones de los números que se han corregido para los cambios de nivel de la series de tiempo y el vector $\sum d_i\, g_i\, \boldsymbol{x}_i$ equivale a los números no corregidos correspondientes. Estos números provienen del gráfico 8.23.
3. Los ponderadores ajustados se vuelven: $w_i = d_i\, g_i\, (1 + \boldsymbol{x}'_i\, \boldsymbol{\lambda})$
   Estos ponderadores dan las siete series que hemos usado como condiciones de calibración y se usan para calcular otros valores de series de tiempo vinculadas.

Gráfico 8.24 Corrección por los cambios de nivel en el Registro de Educación, 1999

Persona	Nivel de escolaridad	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}	x_{6i}	x_{7i}	$d_i\, g_i$	w_i
NIP1	Menos de 9 años	1	0	0	0	0	0	0	1.01693	0.99895
NIP2	9 años obligatorios	0	1	0	0	0	0	0	1.01693	0.91917
NIP3	Media superior: 2 años	0	0	1	0	0	0	0	1.01693	1.03375
NIP4	Media superior: 3 años	0	0	0	1	0	0	0	1.01693	1.08398
NIP5	Universidad: < 3 años	0	0	0	0	1	0	0	1.01693	0.89805
NIP6	Universidad: ≥ 3 años	0	0	0	0	0	1	0	1.01693	1.19574
NIP7	Posgrado	0	0	0	0	0	0	1	1.01693	1.01667
NIP8	Valor perdido	0	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…
NIP6,335,136	Universidad: < 3 años	0	0	0	0	1	0	0	1.01693	0.89805
Total:									**6 335 136**	**6 335 136**

t_x	$\sum d_i g_i x_i$	Diferencia
862 428	877 948	-15 520
942 446	1 042 680	-100 234
1 828 054	1 798 316	29 738
1 109 997	1 041 337	68 660
760 539	861 213	-100 674
789 359	671 318	118 041
42 313	42 324	-11

Las diferencias anteriores corresponden a las correcciones para 1999 en los gráficos 8.21 y 8.23.

De la misma manera, se calculan los ponderadores calibrados para los registros de los años restantes. A continuación mostramos los cálculos para 1998. El ajuste por no respuesta parcial se lleva a cabo con el valor 1.01830 para el ponderador $d_i\, g_i$.

Las siete expresiones $\boldsymbol{t}_x$ son los valores vinculados, mientras que $\sum d_i\, g_i\, \boldsymbol{x}_i$ son los valores de las series de tiempo para 1998 que sólo se corrigieron por no respuesta parcial.

Gráfico 8.25 Corrección por cambios de nivel en el Registro de Educación, 1998

Persona	Nivel de escolaridad	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}	x_{6i}	x_{7i}	$d_i\, g_i$	w_i
NIP1	Menos de 9 años	1	0	0	0	0	0	0	1.01830	1.00299
NIP2	9 años obligatorios	0	1	0	0	0	0	0	1.01830	0.92909
NIP3	Media superior: 2 años	0	0	1	0	0	0	0	1.01830	1.03318
NIP4	Media superior: 3 años	0	0	0	1	0	0	0	1.01830	1.08046
NIP5	Universidad: < 3 años	0	0	0	0	1	0	0	1.01830	0.90862
NIP6	Universidad: ≥ 3 años	0	0	0	0	0	1	0	1.01830	1.18645
NIP7	Posgrado	0	0	0	0	0	0	1	1.01830	1.01804
NIP8	Valor perdido	0	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…
NIP6,335,136	Universidad: < 3 años	0	0	0	0	1	0	0	1.01830	0.90862
Total:									**6 327 060**	**6 327 060**

t_x	$\sum d_i g_i x_i$	Diferencia
915 075	929 043	-13 968
939 556	1 029 766	-90 210
1 858 639	1 831 874	26 764
1 074 101	1 012 306	61 794
750 649	841 256	-90 607
749 600	643 364	106 237
39 440	39 450	-10

Las diferencias corresponden a las correcciones para 1998 en el gráfico 8.21

En el capítulo 9 veremos un método alternativo de corrección para cambios de nivel en las series de tiempo.

CAPÍTULO 9

Estimación con objetos combinados

En este capítulo introducimos el concepto de *objeto combinado*. Los métodos de estimación que usan objetos combinados sirven para reducir los *errores de agregación* y hacer correcciones por cambios de nivel en las series de tiempo. Los errores de agregación pueden ocurrir cuando se integran datos de diferentes registros con distintos tipos de objetos. En la sección 3.3.2 analizamos diferentes tipos de variables, como las de valor único y las de valores múltiples. Una *variable de valor único* sólo puede aceptar *un* valor por cada objeto, mientras que una *variable de valores múltiples* puede aceptar *varios* valores para cierto objeto. Actualmente, las variables de valores múltiples se usan de una manera que origina errores de agregación. En este capítulo presentamos métodos de estimación especiales con ponderadores aplicables a las variables de valores múltiples, que permiten reducir los errores de agregación de las variables de valores múltiples.

9.1 ERRORES DE AGREGACIÓN

En la sección 3.3.4, vimos la agregación y la adjunción como métodos para crear variables derivadas. Agregar una variable cualitativa es una operación que puede generar errores e inconsistencias entre los diferentes registros. En el gráfico siguiente, el número total de empleados es *tres* en el registro 1 y *cinco* en el registro 3. Las sumas salariales por rama industrial en el registro 1 difieren de las de los otros registros.

Gráfico 9.1 Número de empleados y sumas salariales en diferentes registros

Registro 1 – Personas

Persona	Sexo	Suma salarial	1er rama industrial
NIP1	M	450 000	D
NIP2	F	210 000	D
NIP3	M	270 000	A

Registro 2 – Actividades laborales

Em-pleo	Perso-na	Estable-cimiento	Suma salarial	Rama indus-trial	Sexo
Em1	NIP1	Es1	220 000	A	M
Em2	NIP3	Es1	180 000	A	M
Em3	NIP1	Es2	230 000	D	M
Em4	NIP2	Es2	210 000	D	F
Em5	NIP3	Es2	90 000	D	M

Agregación

Agregación

Registro 3 – Establecimiento

Estable-cimiento	Rama industrial	Suma salarial	Núm. de empleados	Proporción de mujeres
Es1	A	400 000	2	0.00
Es2	D	530 000	3	0.33

Las inconsistencias del gráfico 9.1 son ejemplos de un problema general cuando se integran datos de diferentes registros. Incluso cuando todas las variables e identidades de los registros son totalmente correctas, se presentarán estos errores. Como se generan durante el

proceso de integración, reciben el nombre de *errores de integración*. En este capítulo veremos un tipo específico de errores de integración a los que llamamos *errores de agregación*.

Cuando agregamos valores de variables de *muchos objetos a uno*, ocurrirá un error de agregación si la variable es cualitativa. Se presentará el mismo problema si un objeto puede ocurrir varias veces en el registro, pero con diferentes características, por ejemplo, cuando un objeto ha cambiado durante el periodo de referencia del registro. El problema con los errores de agregación puede deberse a:

– objetos que ocurren varias veces en el registro;
– relaciones muchos a uno;
– variables de valores múltiples.

Objetos que ocurren varias veces: registros por año calendario

En un registro por año calendario, los objetos que cambian durante el año ocurren varias veces. Las personas se mudan o cambian de estado civil; los hogares se modifican; las empresas se dividen o fusionan o cambian de rama industrial. Algunos objetos pueden cambiar muchas veces en un mismo año y todos esos cambios generan variables de valores múltiples. El ejemplo del gráfico 7.8 continúa en el gráfico siguiente, donde las personas se mudaron durante 2005.

Gráfico 9.2 Registro por año calendario para la población de personas durante 2005

Persona	Domicilio	Municipio	Desde aaaammdd	Hasta aaaammdd	Ponderador = tiempo en el domicilio (años)
NIP1	Domicilio 1	1	20050101	20050517	136/365 = 0.37
NIP1	Domicilio 2	2	20050518	20051231	229/365 = 0.63
NIP2	Domicilio 3	1	20050101	20051231	365/365 = 1.00
NIP3	Domicilio 4	2	20050101	20050314	73/365 = 0.20
NIP3	Domicilio 5	1	20050315	20050925	194/365 = 0.53
NIP3	Domicilio 6	2	20050926	20051231	98/365 = 0.27
NIP4	Domicilio 7	2	20050101	20050605	156/365 = 0.43
NIP4	Domicilio 8	1	20050606	20051231	209/365 = 0.57

El domicilio y el municipio son *variables de valores múltiples* en este ejemplo. El tiempo se puede usar para generar ponderadores para cada combinación de persona y domicilio. El registro contiene *cuatro* personas, pero son *ocho* combinaciones de persona y domicilio. Como todas las personas viven todo el año, los ponderadores de cada persona suman 1. Con estos ponderadores se puede estimar la distribución de frecuencias de personas por municipio.

Gráfico 9.3 Población promedio, 2005

Municipio	Frecuencia absoluta	Frecuencia relativa
1	0.37 + 1.00 + 0.53 + 0.57 = 2.47	62%
2	0.63 + 0.20 + 0.27 + 0.43 = 1.53	38%
Total	4.00	100%

Los registros por año calendario constituyen una clase importante de registros con problemas de estimación ocasionalmente complejos.

Relaciones muchos a uno

En el ejemplo anterior, hay *un* registro en el que algunos objetos aparecen varias veces, lo que da origen a variables de valores múltiples. En otros casos se emparejan *dos* registros con diferentes tipos de objetos. Cuando hay muchas relaciones muchos a uno, se pueden crear variables de valores múltiples al integrar estos registros y agregar variables cualitativas. Si una persona tiene más de un empleo, ¿cómo se debe agregar la información sobre estos empleos a la información sobre la persona? El gráfico 9.1 puede servir como ejemplo. A continuación el registro 2 está ordenado por NIP.

Gráfico 9.4 Número de empleados y sumas salariales en diferentes registros

Registro 1 – Personas

Persona	Sexo	Suma salarial	1er rama industrial
NIP1	M	450 000	D
NIP2	F	210 000	D
NIP3	M	270 000	A

Agregación

Registro 2 – Actividades laborales

Empleo	Persona	Establecimiento	Suma salarial	Rama industrial	Sexo
Em1	NIP1	Es1	220 000	A	M
Em3	NIP1	Es2	230 000	D	M
Em4	NIP2	Es2	210 000	D	F
Em2	NIP3	Es1	180 000	A	M
Em5	NIP3	Es2	90 000	D	M

En el registro 2, la rama industrial es una variable de valor único que describe una característica del tipo de objeto *empleo o actividad*. NIP1 y NIP3 tienen dos empleos, ambas personas trabajan en un establecimiento de la rama industrial A y también en un establecimiento de la rama industrial D. La manera tradicional de crear una variable *rama industrial* para personas consiste en usar información sobre *un solo* empleo para cada persona: el empleo más importante. En esta situación, cuando la mayoría de las personas tienen un empleo, pero algunas tienen más de uno aunque sea un número de empleos limitado, es mejor definir las variables *establecimiento* y *rama industrial para personas* como variables de valores múltiples. En el Registro 2 del gráfico 9.4, estas variables son de valor único para *empleos*, mientras que en el gráfico 9.5 se han convertido en variables de valores múltiples para *personas*. Las sumas salariales para cada persona se usan para crear ponderadores; los ponderadores para cada persona deben sumar 1.

Gráfico 9.5 Número de empleados y sumas salariales en diferentes registros

Persona	Sexo	Establecimiento	Suma salarial	Rama industrial	Ponderador
NIP1	M	Es1	220 000	A	22/45 = 0.49
NIP1	M	Es2	230 000	D	23/45 = 0.51
NIP2	F	Es2	210 000	D	21/21 = 1.00
NIP3	M	Es1	180 000	A	18/27 = 0.67
NIP3	M	Es2	90 000	D	9/27 = 0.33

La matriz de datos del gráfico 9.5 se puede usar para estimar tablas que incluyan personas por rama industrial.

Gráfico 9.6 Empleados por rama industrial

Rama industrial	**Número de empleados**
A	0.49 + 0.67 = 1.16
D	0.51 + 1.00 + 0.33 = 1.84
Total	3.00

Los dos ejemplos de esta sección muestran que pueden surgir variables de valores múltiples por diferentes motivos. En la siguiente sección, nos ocuparemos de las variables de valores múltiples en los sistemas de registros y propondremos métodos para obtener estimaciones consistentes. Las inconsistencias del gráfico 9.1 se vuelven innecesarias aplicando estos métodos.

9.2 MÉTODOS DE ESTIMACIÓN PARA VARIABLES DE VALORES MÚLTIPLES

En el Registro de Educación, que es un registro sobre personas, se creó la variable *nivel máximo de escolaridad*. Esta variable es de valores múltiples porque algunas personas tienen dos o más títulos del mismo nivel. En el Registro Empresarial se crea la variable *rama industrial*, que también es una variable de valores múltiples de gran importancia. Éstos son ejemplos de variables con valores múltiples en la fuente original. Como mencionamos en la sección anterior, las variables de valores múltiples también se crean en el sistema de registros.

Aunque difíciles de manejar, las variables de valores múltiples son frecuentes e importantes dentro del sistema de registros. Las variables de valores múltiples de los registros estadísticos también se usan en censos y encuestas por muestreo, lo que significa que el problema de su manejo también afectará estos tipos de encuesta.

Por lo general, este problema se "resuelve" de manera drástica: la variable de valores múltiples se convierte en una variable de valor único usando sólo "el valor más importante" para cada objeto. Si, por ejemplo, se va a describir la distribución de personas por ocupación, se subestimarán las ocupaciones comunes como actividades secundarias. Como se descarta una parte de la información ocupacional, entonces las estimaciones tendrán problemas de calidad de una magnitud desconocida como consecuencia de los errores de agregación.

Empezaremos con un ejemplo sencillo para mostrar los principios fundamentales del manejo que, en nuestra opinión, se debe dar a las variables de valores múltiples. Después abordaremos algunas situaciones más complejas que se presentan cuando se ponen en práctica los principios.

9.2.1 Ocupación en los registros de Actividades y de Ocupaciones

En esta sección, el problema de estimación se puede plantear de la siguiente manera: ¿cómo debe estimarse la distribución de frecuencias de las diferentes ocupaciones?

Principio 1:
¿Qué cálculos deben efectuarse? El problema de estimación siempre se debe especificar antes de iniciar los cálculos. Esto se ilustra a continuación comparando tres formas de definirlo y resolverlo.

La matriz de datos del gráfico 9.7 muestra la ocupación y el código ocupacional (CIUO) de seis personas, dos de las cuales tienen más de una ocupación. El objeto de la matriz es *empleo*, que es un *objeto relacional* identificado mediante el número de identificación personal y la identidad de unidad jurídica para cada empresa. La variable *jornada*, es decir, la jornada de trabajo (tiempo completo o tiempo parcial), se presenta como porcentaje de un empleo de tiempo completo. Esta variable se toma del registro salarial. Supongamos que la matriz de datos contiene todas las actividades ocupacionales en una región pequeña, ¿cómo debemos estimar la distribución de *personas* en diferentes ocupaciones? Ése es nuestro primer problema de estimación.

Gráfico 9.7 Registro de Empleo con datos ocupacionales

Id. de empleo	Persona	Unidad jurídica	Ocupación	CIUO	Jornada
Em1	NIP1	UJ1	Estadístico	2211	100
Em2	NIP1	UJ2	Agricultor	6111	15
Em3	NIP1	UJ3	Político	1110	10
Em4	NIP2	UJ4	Camillero de hospital	5132	30
Em5	NIP2	UJ5	Trabajador de limpieza	9122	20
Em6	NIP3	UJ6	Dependiente de tienda	5221	10
Em7	NIP4	UJ6	Dependiente de tienda	5221	50
Em8	NIP5	UJ6	Dependiente de tienda	5221	20
Em9	NIP6	UJ6	Dependiente de tienda	5221	100
Total					

El enfoque tradicional es que cada persona tiene *una* sola ocupación, es decir, la *ocupación principal*. Esto significa que se descarta información sobre quienes tienen varias ocupaciones, pues sólo se incluye la ocupación con más horas de trabajo.

Entonces generamos la nueva matriz de datos (gráfico 9.8), en la que el objeto es *persona* y la distribución por ocupación se obtiene sumando el número de personas en cada ocupación.

Gráfico 9.8 Registro tradicional sobre personas con información ocupacional

Persona	Unidad jurídica	Ocupación principal	CIUO	Jornada	Ponderador alternativo 1
NIP1	UJ1	Estadístico	2211	100	1
NIP2	UJ4	Camillero de hospital	5132	30	1
NIP3	UJ6	Dependiente de tienda	5221	10	1
NIP4	UJ6	Dependiente de tienda	5221	50	1
NIP5	UJ6	Dependiente de tienda	5221	20	1
NIP6	UJ6	Dependiente de tienda	5221	100	1
Total					**6**

El gráfico 9.9 contiene la distribución ocupacional estimada. De acuerdo con la tradición en las estadísticas sobre personas, cada persona tendrá la misma ponderación independientemente de que trabaje 100 o 10% de un empleo de tiempo completo.

Gráfico 9.9 Personas empleadas por ocupación, alternativa tradicional 1

Ocupación principal	CIUO	Número	Porcentaje
Estadístico	2211	1	16.7
Camillero de hospital	5132	1	16.7
Dependiente de tienda	5221	4	66.7
Total		**6**	**100.0**

Este ejemplo muestra que se subestiman ocupaciones secundarias comunes, como las de *político* y *agricultor* —a menudo ejercidas de manera paralela a la ocupación principal—. En cambio, podrían hacerse estimaciones para variables de valores múltiples de manera que no se descarte ninguna información. Esto será posible si las estimaciones se basan en una matriz de datos con "objetos combinados".

Principio 2:
El principio básico consiste en crear una matriz de datos de modo que *cada combinación de objetos y valores de la variable de valores múltiples corresponda a un objeto* en la nueva matriz de datos. Los objetos, o renglones, en estas matrices de datos se denominan *objetos combinados*.

La matriz de datos del gráfico siguiente se formó de esta manera: las seis personas del registro sobre personas del gráfico 9.8 dieron origen a nueve objetos combinados.

Gráfico 9.10 Registro sobre personas con datos ocupacionales

Objeto combinado	Persona	Ocupación	Jornada	Ponderador alternativo 1	Ponderador alternativo 2
1	NIP1	Estadístico	100	1	0.80
2	NIP1	Agricultor	15	0	0.12
3	NIP1	Político	10	0	0.08
4	NIP2	Camillero de hospital	30	1	0.60
5	NIP2	Trabajador de limpieza	20	0	0.40
6	NIP3	Dependiente de tienda	10	1	1.00
7	NIP4	Dependiente de tienda	50	1	1.00
8	NIP5	Dependiente de tienda	20	1	1.00
9	NIP6	Dependiente de tienda	100	1	1.00
Total				**6**	**6.00**

Los renglones de la matriz son todas las combinaciones de *persona·ocupación*. Por ejemplo, la persona *NIP1*, que tiene tres ocupaciones, aparece en tres renglones. Los ponderadores de acuerdo con la alternativa 2 se calcularon con la variable *jornada*, de modo que 0.80 = 100/(100 + 15 + 10), etc. Los ponderadores para cada persona suman 1 en las alternativas 1 y 2, y la suma de todos los ponderadores de ambas alternativas es 6, es decir, el número total de personas. En el gráfico 9.10 se ilustran varios principios generales:

Principio 3:
La suma de ponderadores para *una persona* (el tipo de objeto al que se refiere el problema de estimación y que fue el punto de partida cuando se formaron los objetos combinados) siempre deberá ser 1.

Principio 4:
Del principio 3 se desprende que la suma de todos los ponderadores es igual al número total de objetos (el tipo de objeto al que se refiere el problema de estimación).

En el siguiente gráfico, se calcula la distribución de frecuencias de personas por ocupación con ponderadores de acuerdo con las alternativas 1 y 2 del gráfico 9.10.

Gráfico 9.11 Personas empleadas por ocupación de acuerdo con las dos alternativas

		Alternativa 1		Alternativa 2	
Ocupación	**CIUO**	**Núm.**	**Porcentaje**	**Núm.**	**Porcentaje**
Político	1110	0.00	0.0	0.08	1.3
Estadístico	2211	1.00	16.7	0.80	13.3
Camillero de hospital	5132	1.00	16.7	0.60	10.0
Dependiente de tienda	5221	4.00	66.7	4.00	66.7
Agricultor	6111	0.00	0.0	0.12	2.0
Trabajador de limpieza	9122	0.00	0.0	0.40	6.7
Total		**6.00**	**100.0**	**6.00**	**100.0**

El número de personas empleadas (Núm.) por ocupación se estima sumando los ponderadores de cada ocupación. El ponderador alternativo 1 del gráfico 9.10 se suma en la alternativa 1 y el ponderador alternativo 2 se suma en la alternativa 2.

Como algunas de las ocupaciones tienen el ponderador 0 en la alternativa 1, según el método tradicional de cálculo, se producen errores de agregación y las estimaciones conforme a esa alternativa están distorsionadas, ya que se sobrestima la frecuencia de ciertas ocupacio-

nes y se subestima la de otras. En cambio, las estimaciones conforme a la alternativa 2 utilizan toda la información sobre las ocupaciones en la variable de valores múltiples.

Los ponderadores de la alternativa 2 usan la variable *jornada*, que se encuentra en el Registro Salarial sueco. Para las posiciones no incluidas en ese registro, se deben calcular ponderadores a partir de otra información. El Registro de Estados de Resultados contiene el *salario bruto anual* para todos los empleos, por lo que siempre será factible usarlo. Los ponderadores calculados a partir del salario bruto anual son un tanto distintos de los calculados a partir de la *jornada.* Para obtener buenas estimaciones, se debe elegir entre diferentes *variables generadoras de ponderadores* y es importante seleccionar una variable que sea tanto relevante como funcional.

Los ponderadores de acuerdo con las alternativas 1 y 2 se basan en la variable *jornada*, pero también podrían basarse en otras variables. Los ponderadores usados en la práctica pueden diferir en mayor o menor grado de los ponderadores ideales; para ciertas personas, los ponderadores para una ocupación pueden ser demasiado grandes, mientras que para otras, los ponderadores para esa misma ocupación pueden ser demasiado pequeños. Los errores se pueden compensar en parte cuando se hace la distribución general de ocupaciones. La medición de calidad relevante podría ser una medición de qué tan cerca se encuentra la distribución estimada de la distribución que se calcularía con los ponderadores ideales.

Principio 5:
Es mejor utilizar los ponderadores aproximados que los ponderadores completamente incorrectos.

En las alternativas 1 y 2, el problema de estimación consiste en describir la distribución de *personas* por ocupación. Otra posibilidad, la *alternativa 3*, es distribuir la *jornada de trabajo* por ocupación. La jornada o el volumen de trabajo pueden describirse según la cantidad de actividad ocupacional recalculada en relación con personas empleadas de tiempo completo. Este método de cálculo es común en las estadísticas económicas, donde se suelen medir volúmenes y no personas.

La persona *NIP1* tiene tres ocupaciones, una de tiempo completo y otras dos que corresponden a 15 y 10% de un empleo de tiempo completo. La matriz con las seis personas representa 3.55 *empleos de tiempo completo*. El objeto de la matriz del gráfico 9.12 es *empleo* y la variable *ocupación* es una variable de valor único: cada empleo corresponde a una sola ocupación. La distribución de empleos de tiempo completo por ocupación se obtiene sumando la variable *ponderador alternativo 3 (= jornada /100* en el gráfico 9.10*)* para las diferentes ocupaciones.

Gráfico 9.12 Registro sobre personas con datos ocupacionales

Persona	Ocupación	Ponderador alternativo 1	Ponderador alternativo 2	Ponderador alternativo 3
NIP1	Estadístico	1	0.80	1.00
NIP1	Agricultor	0	0.12	0.15
NIP1	Político	0	0.08	0.10
NIP2	Camillero de hospital	1	0.60	0.30
NIP2	Trabajador de limpieza	0	0.40	0.20
NIP3	Dependiente de tienda	1	1.00	0.10
NIP4	Dependiente de tienda	1	1.00	0.50
NIP5	Dependiente de tienda	1	1.00	0.20
NIP6	Dependiente de tienda	1	1.00	1.00
Total		**6**	**6.00**	**3.55**

Gráfico 9.13 Personas y empleos de tiempo completo por ocupación, tres alternativas

		Alternativa 1		Alternativa 2		Alternativa 3	
Ocupación	**CIUO**	**Núm.**	**%**	**Núm.**	**%**	**Núm.**	**%**
Político	1110	0.00	0.0	0.08	1.3	0.10	2.8
Estadístico	2211	1.00	16.7	0.80	13.3	1.00	28.2
Camillero de hospital	5132	1.00	16.7	0.60	10.0	0.30	8.5
Dependiente de tienda	5221	4.00	66.7	4.00	66.7	1.80	50.7
Agricultor	6111	0.00	0.0	0.12	2.0	0.15	4.2
Trabajador de limpieza	9122	0.00	0.0	0.40	6.7	0.20	5.6
Total		**6.00**	**100.0**	**6.00**	**100.0**	**3.55**	**100.0**

Las alternativas 1 y 2 se relacionan con el mismo problema de estimación, *personas* distribuidas por ocupación, pero se basan en métodos de estimación distintos que usan ponderadores distintos.

La alternativa 3 se relaciona con otro problema de estimación, *jornada de trabajo* distribuida por ocupación, donde el enfoque se da en el volumen de trabajo, no en las personas.

9.2.2 Clasificación industrial en el Registro Empresarial

La clasificación industrial es otra variable de valores múltiples importante. Se crea en el Registro Empresarial y se usa en muchos registros dentro del sistema de registros. En este caso también es común seleccionar la "clasificación industrial más importante" y descartar información sobre otras clasificaciones industriales para los establecimientos o las unidades empresariales a las que se refieren las estadísticas. Esto provoca errores de agregación en las estadísticas económicas.

El Registro Empresarial de la Oficina de Estadística de Suecia contiene información sobre todas las ramas industriales en las que participa una empresa. También hay detalles sobre la proporción de empresas que operan en cada rama. El código de clasificación industrial y la participación en cada rama industrial son de buena calidad cuando se relacionan con empresas manufactureras. El método para elegir la rama más importante de la industria puede ser problemático en cuanto a la presentación de estadísticas industriales y las series de tiempo. Por ejemplo, si 51% de las actividades de una empresa grande en el año 1 entran dentro de cierta clasificación industrial, pero sólo 49% corresponde a esa misma clasificación en el año 2, esa pequeña variación puede originar cambios de nivel considerables en muchas series de tiempo, en las que, por ejemplo, todos los empleados que trabajan en esa empresa aparentemente cambian de clasificación industrial del año 1 al 2.

Estos problemas pueden ser aun más graves en las estadísticas regionales, pues un establecimiento puede ser dominante, lo cual significa que un cambio en la clasificación industrial acarreará cambios de nivel en las series de tiempo dentro de las series regionales.

Los cambios menores son más problemáticos todavía por ser más difíciles de detectar y en muchos casos se malinterpretarán como cambios reales en la economía. La metodología presentada en la sección anterior, respecto a la ocupación, permite evitar estos problemas de calidad.

En el gráfico 9.14, se muestra la clasificación industrial y el número de empleados para tres establecimientos. *El problema de estimación tiene que ver con el número de empleados por clasificación industrial*. La información utilizada en el método tradicional aparece en las celdas sombreadas, mientras que las celdas no sombreadas contienen información disponible, pero no utilizada.

Gráfico 9.14a Registro Empresarial, año 1: matriz de datos para establecimientos

Establecimiento	Rama industrial 1 [1]	%	Rama industrial 2	%	Rama industrial 3	%	Núm. de empleados
Es1	DJ	100					218
Es2	DH	51	DJ	49			293
Es3	DJ	40	DH	30	DK	30	156

Gráfico 9.14b Registro Empresarial, año 2: matriz de datos para establecimientos

Establecimiento	Rama industrial 1	%	Rama industrial 2	%	Rama industrial 3	%	Núm. de empleados
Es1	DJ	100					221
Es2	DJ	52	DH	48			314
Es3	DJ	36	DH	34	DK	30	143

Gráfico 9.14c Número de empleados por rama industrial, estimaciones tradicionales

Rama industrial	Año 1	Año 2
DH	293	0
DJ	374	678
DK	0	0
Total	**667**	**678**

El número de empleados se ordena por rama industrial principal. Ésa es la manera más común de presentar series de tiempo con base en la clasificación industrial del Registro Empresarial.

Esto ocasiona cambios abruptos en las series que presentamos aquí.

En la sección 9.2.7 damos un ejemplo que ilustra los problemas con las series de tiempo, que pueden derivarse de los métodos de estimación tradicionales. En ese ejemplo, basado en datos reales, hay una grave alteración en las series de tiempo.

En los gráficos 9.14a y 9.14b, los porcentajes muestran la proporción de cada rama industrial, que es una medida del tamaño de cada rama en cada establecimiento. El criterio para la medición del tamaño puede ser la facturación, el número de empleados o algún otro.

Para empezar, damos por sentado que los gráficos porcentuales se basan en el número de empleados y luego mostramos cómo los ponderadores establecidos de acuerdo con una medición de tamaño específica se convierten en ponderadores definidos conforme a otra medición de tamaño.

Según el *principio 2* enunciado en la sección anterior, se crea una nueva matriz de datos (véase gráfico 9.15) que contiene objetos combinados, de modo que cada combinación de objetos y valores para la variable de valores múltiples corresponda a un renglón de la nueva matriz de datos.

En esta matriz de datos, cada renglón es una combinación de establecimiento y clasificación industrial. En vez de una matriz de *tres* renglones correspondientes a tres establecimientos, obtenemos una matriz de *seis* renglones correspondientes a todas las combinaciones *establecimiento* • *rama industrial* para cada año. Ahora podemos estimar el número de empleados en cada rama industrial mediante la fórmula (4):

$$\hat{Y} = \sum_{i=1}^{R} w_i y_i \qquad \text{En el gráfico 9.15 se calculó } w_i y_i \text{ para cada renglón} \qquad (4)$$

[1] Fabricación de : DM = Vehículos de motor; DK = Maquinaria; DB = Textiles; DA = Productos alimenticios; DH = Industria de la transformación del caucho y materias plásticas; DJ = Metalurgia, fabricación de productos metálicos; DK = Industria de la construcción de maquinaria y equipo mecánico.

En la sección anterior cuando hablamos de ocupación, la variable y_i = 1 para todos los objetos combinados. La fórmula (4) nos indica que se suman los ponderadores w_i para todas las celdas de la tabla del gráfico 9.13.

Gráfico 9.15 Matriz de datos con objetos combinados: establecimiento • rama industrial

Año 1				
Establecimiento	Rama industrial	Ponderador, w_i	Núm. de empleados, y_i	$w_i\, y_i$
Es1	DJ	1.00	218	218
Es2	DH	0.51	293	149.43
Es2	DJ	0.49	293	143.57
Es3	DJ	0.40	156	62.4
Es3	DH	0.30	156	46.8
Es3	DK	0.30	156	46.8
Total		**3.00**		**667**

Año 2				
Establecimiento	Rama industrial	Ponderador, w_i	Núm. de empleados, y_i	$w_i\, y_i$
Es1	DJ	1.00	221	221
Es2	DH	0.48	314	150.72
Es2	DJ	0.52	314	163.28
Es3	DJ	0.36	143	51.48
Es3	DH	0.34	143	48.62
Es3	DK	0.30	143	42.9
Total		**3.00**		**678**

Los ponderadores w_i suman 3, porque seguimos hablando de tres establecimientos. Las sumas de los productos $w_i\, y_i$ darán los mismos totales que antes:

El número total de empleados es un total dado que no debe cambiar cuando se introducen ponderadores. Estos últimos sólo afectarán la distribución de los empleados entre las diferentes ramas industriales.

Éste es un ejemplo de un principio de aplicación general.

Gráfico 9.16 Número de empleados por rama industrial, estimado con objetos combinados

Rama industrial	Año 1	Año 2
DH	196.23	199.34
DJ	423.97	435.76
DK	46.80	42.90
Total	**667.00**	**678.00**

Las series de tiempo del gráfico 9.16 se calcularon con los ponderadores w_i.

Estas series tienen una mayor calidad que las del gráfico 9.14c, con modificaciones relevantes y sin cambios de nivel.

9.2.3 Transformación de los ponderadores

Los ponderadores para las diferentes clasificaciones industriales del Registro Empresarial sueco se basan sobre todo en la facturación. Por lo tanto, se deben convertir cuando se calculan estimaciones para otras variables.

Ejemplo: Cuando se estima el número de empleados, los ponderadores basados en la facturación se deben convertir a ponderadores adecuados para el número de empleados. Para esta conversión se utiliza un modelo que describe la relación entre la facturación y el número de empleados; el modelo, por su parte, se basa en estadísticas apropiadas que describen el empleo y la facturación para establecimientos de una rama industrial específica. En el gráfico 9.17 se muestra esta conversión para el establecimiento *Es3.* El ponderador derivado de la facturación se multiplica por el número de empleados según la facturación. Estos valores se deben dividir entre una constante, de modo que el total sea uno para cada establecimiento.

Gráfico 9.17 Conversión de ponderadores

Registro de establecimientos			Datos agregados		Registro de establecimientos		
Año 1		Ponderadores	Modelos para diferentes ramas industriales		Ponderadores convertidos con base en el modelo		
Esta-bleci-miento	Rama industrial	Con base en la facturación	$\frac{\text{Empleados}}{\text{Facturación (millones de coronas suecas)}}$		Adaptado para la estimación del número de empleados		
Es3	DJ	0.4	DJ	0.5	Es3	DJ	$\frac{0.4 \cdot 0.5}{(0.4 \cdot 0.5 + 0.3 \cdot 0.6 + 0.3 \cdot 0.7)} = \mathbf{0.34}$
Es3	DH	0.3	DH	0.6	Es3	DH	$\frac{0.3 \cdot 0.6}{(0.4 \cdot 0.5 + 0.3 \cdot 0.6 + 0.3 \cdot 0.7)} = \mathbf{0.30}$
Es3	DK	0.3	DK	0.7	Es3	DK	$\frac{0.3 \cdot 0.7}{(0.4 \cdot 0.5 + 0.3 \cdot 0.6 + 0.3 \cdot 0.7)} = \mathbf{0.36}$

De acuerdo con el principio 3, la suma de los ponderadores para *un establecimiento* (el tipo de objeto al que se refiere el problema de estimación) siempre debe ser 1.

El ponderador convertido para la participación del establecimiento perteneciente a la industria siderúrgica intensiva en capital, DJ, es menor que el ponderador original, que se basó en la facturación (0.34 en comparación con 0.4). El ejemplo también muestra que, cuando se usa la facturación, la industria más importante es DJ, pero si se considera el número de empleados, la más importante sería DK.

Esto demuestra una vez más que el principio de sólo usar el valor "más importante" de una variable de valores múltiples puede causar problemas. Además, cuando los registros contienen variables de valores múltiples, se deben calcular ponderadores que se adapten a diferentes problemas de estimación.

9.2.4 Importación de una gran cantidad de variables de valores múltiples

El Registro de Empleo importa una gran cantidad de variables de diversos registros. La población se crea a partir del Registro de Población al 31 de diciembre, de donde también se toman variables de valor único como edad y sexo. Las variables de valores múltiples *nivel de escolaridad* y *ocupación* provienen de los registros de Educación y Ocupación, respectivamente.

El Registro de Actividades contiene el objeto *empleo*, que es una relación entre una persona y un establecimiento. La clasificación industrial para el establecimiento se importa del Registro Empresarial al Registro de Actividades. Por último, la identidad del establecimiento, junto con su clasificación industrial, se exporta del Registro de Actividades al Registro de Empleo. Tanto la identidad del establecimiento como la clasificación industrial son variables de valores múltiples en el Registro de Empleo.

En el Registro de Empleo se crea una variable derivada para mostrar si una persona tuvo un empleo remunerado en noviembre (llamada *EmpNov*). En la sección 6.2.3 describimos el método con el que se crea esta variable.

Con la metodología de estimación tradicional, sólo se usan los datos relativos al valor "más importante" de todas estas variables de valores múltiples (sombreados en las matrices de datos del gráfico 9.18).

Primero mostramos todos los registros utilizados con los datos vinculados a una persona específica, *NIP10*, y cómo funciona la estimación tradicional. Luego explicamos cómo usar toda la información con objetos combinados para las variables de valores múltiples.

1. Metodología tradicional: sólo se utiliza el valor más importante

Gráfico 9.18a Registro de Población

Persona	Sexo	Edad
NIP10	F	32

Gráfico 9.18b Registro de Educación

Persona	Nivel de escolaridad 1	Puntos 1	Nivel de escolaridad 2	Puntos 2
NIP10	NE1	180	NI2	120

NIP10 tiene dos títulos del mismo nivel en diferentes campos; Nivel de escolaridad 2 es el más reciente.

Gráfico 9.18c Registro de Actividades, con jornada de trabajo en noviembre

Persona	Establecimiento	Jornada
NIP10	Es11	80%
NIP10	Es12	20%

Tradicionalmente, sólo se usa el establecimiento de la actividad principal.

Gráfico 9.18d Registro Ocupacional

Persona	Establecimiento	Ocupación
NIP10	Es11	Oc1
NIP10	Es12	Oc2

Tradicionalmente, sólo se usa la ocupación de la actividad principal.

Gráfico 9.18e Registro Empresarial

Establecimiento	Rama industrial	Ponderador 1	Rama industrial	Ponderador 2
Es11	DH	70%	DJ	30%
Es12	DK	100%		

Los registros de Actividades y Empresarial se emparejan usando la identidad del establecimiento como variable de vinculación. Se importa la mayor *rama industrial* al Registro de Actividades, que después se empareja con el Registro Ocupacional usando el número de identificación personal y la identidad del establecimiento como variables de vinculación. La *ocupación* se importa al Registro de Actividades.

Gráfico 9.18f Registro de Actividades, se importan la rama industrial y la ocupación

Persona	Establecimiento	Rama industrial	Ocupación
NIP10	Es11	DH	Oc1
NIP10	Es12	DK	Oc2

Tradicionalmente, sólo se usan el establecimiento y la ocupación de la actividad principal.

El Registro de Empleo se crea en tres etapas:

– Se establece un nuevo registro con personas de 16 años o más de edad, con las variables *sexo* y *edad* del Registro de Población.

– Los valores más importantes de las variables de valores múltiples para *nivel de escolaridad*, *ocupación*, *establecimiento* y *clasificación industrial* se importan de los diferentes registros.

– Se crea la variable *EmpNov* para las personas que tenían un empleo remunerado en noviembre.

Aquí vemos la parte del Registro de Empleo completado en relación con la persona *NIP10*:

Gráfico 9.18g Registro de Empleo, datos para la persona NIP10

Persona	Sexo	Edad	Nivel de escolaridad	Ocupación	Establecimiento	Rama industrial	EmpNov
NIP10	F	32	NE2	Oc1	Es11	DH	Sí

2. Metodología con objetos combinados, cuando se usa toda la información

Gráfico 9.19a Registro de Población

Persona	Sexo	Edad
NIP10	F	32

Gráfico 9.19b Registro de Educación

Persona	Nivel de escolaridad	w_{NE}
NIP10	NE1	0.6
NIP10	NE2	0.4

Los ponderadores para nivel de escolaridad se crean usando la duración del programa educativo expresada como "puntos" de escolaridad.

Gráfico 9.19c Registro de Actividades, con duración de la jornada de trabajo en noviembre

Persona	Establecimiento	w_{Es}
NIP10	Es11	0.8
NIP10	Es12	0.2

Para el objeto *persona* se crean ponderadores para la variable de valores múltiples *establecimiento* usando la variable *jornada laboral*.

Gráfico 9.19d Registro Ocupacional

Persona	Establecimiento	Ocupación
NIP10	Es11	Oc1
NIP10	Es12	Oc2

La ocupación está vinculada con la relación entre *persona* y *establecimiento*; el ponderador para *ocupación* es el mismo que para establecimiento.

Gráfico 9.19e Registro Empresarial

Establecimiento	Rama industrial	w_{Ind}
Es11	DH	0.7
Es11	DJ	0.3
Es12	DK	1.0

Con base en información del Registro Empresarial, se crea un registro con el objeto combinado *establecimiento · rama industrial* y los ponderadores para diferentes ramas industriales.

Se crea un registro de personas empleadas con objetos combinados que son combinaciones del tipo de objeto relevante *persona* y *todos* los valores de *todas* las variables de valores múltiples, de acuerdo con el principio antes mencionado. (*Principio 2:* Crear una matriz de datos con objetos combinados de modo que *cada combinación de objetos y valores de la variable de valores múltiples corresponda a un renglón* en la nueva matriz de datos).

Como resultado de los emparejamientos, se crea una matriz de datos. En el gráfico 9.19f vemos la parte de la matriz de datos que se refiere a la persona *NIP10*.

Gráfico 9.19f Objetos combinados: persona • nivel de escolaridad • establecimiento • rama industrial

Persona	Sexo	Edad	Nivel de escola-ridad	Estable-cimiento	Ocupa-ción	Rama industrial	EmpNov	w_{NE}	w_{Es}	w_{Ind}	$w_{ObjComb}$
NIP10	F	32	NE1	Es11	Oc1	DH	Sí	0.6	0.8	0.7	0.336
NIP10	F	32	NE1	Es11	Oc1	DJ	Sí	0.6	0.8	0.3	0.144
NIP10	F	32	NE1	Es12	Oc2	DK	Sí	0.6	0.2	1.0	0.120
NIP10	F	32	NE2	Es11	Oc1	DH	Sí	0.4	0.8	0.7	0.224
NIP10	F	32	NE2	Es11	Oc1	DJ	Sí	0.4	0.8	0.3	0.096
NIP10	F	32	NE2	Es12	Oc2	DK	Sí	0.4	0.2	1.0	0.080
Total											**1.000**

Los cálculos deben referirse al tipo de objeto *persona*. Las variables *nivel de escolaridad*, *establecimiento* y *ocupación* son variables de valor múltiple para el tipo de objeto *persona*. *Rama industrial* es una variable de valores múltiples para el tipo de objeto *establecimiento*. Cada combinación de *persona* y *establecimiento* corresponde a una sola *ocupación*, por lo que no es necesario un ponderador adicional para la variable *ocupación*. En este caso, *ocupación* es una variable de valor único para cada combinación de *persona* y *establecimiento*.

Principio 6:
Cuando hay una gran cantidad de variables de valores múltiples en la misma matriz de datos, todos los ponderadores correspondientes a esas variables se multiplican a fin de obtener los ponderadores que se usarán para la estimación.

En el gráfico 9.19f aparece la persona *NIP10* dividida en seis objetos combinados o renglones. Los ponderadores $w_{ObjComb}$ para estos seis objetos combinados deben sumar 1 y se forman por multiplicación, por ejemplo, 0.6 • 0.8 • 0.7 = 0.336

3. La estimación tradicional comparada con la estimación a partir de objetos combinados

Con las variables edad, sexo, nivel de escolaridad, ocupación y rama industrial, se pueden formar tablas de frecuencias para la variable *EmpNov* (persona con empleo remunerado en noviembre). Comenzando con la matriz de datos del gráfico 9.19f, ahora mostramos la manera en que la persona *NIP10 contribuye a las estimaciones* de las frecuencias en las diferentes celdas de la tabla.

En el gráfico 9.20, las frecuencias de las tablas de la izquierda están estimadas de manera tradicional y las de las tablas de la derecha mediante la suma de los ponderadores $w_{ObjComb}$ en cada celda. Hay muchas diferencias entre los métodos de estimación. En el método tradicional, por ejemplo, gran parte de la información para las variables de valores múltiples se descarta, lo que ocasiona errores de agregación.

Gráfico 9.20a Número de personas con empleo remunerado en noviembre, por edad y sexo

Estimación tradicional

Edad	F	M	Total
20–49	1	0	**1**
50–64	0	0	**0**
65–	0	0	**0**
Total	**1**	**0**	

Las estimaciones para las variables de valor único, como sexo y edad, no se ven afectadas por los ponderadores calculados para las variables de valores múltiples.

Estimación con ponderadores

Edad	F	M	Total
20–49	1	0	**1**
50–64	0	0	**0**
65–	0	0	**0**
Total	**1**	**0**	**1**

Gráfico 9.20b Número de personas con empleo remunerado en noviembre, por ocupación

Estimación tradicional

Ocupación	Número
Oc1	1
Oc2	0
Total	1

Estimación con ponderadores:
0.8 =
0.336 + 0.144 + 0.224 +0.096

Estimación con ponderadores

Ocupación	Número
Oc1	0.8
Oc2	0.2
Total	1

Gráfico 9.20c Núm. de personas con empleo remunerado en noviembre, por nivel de escolaridad

Estimación tradicional

Nivel de escolaridad	Número
NE1	0
NE2	1
Total	1

Estimación con ponderadores:
0.6 = 0.336 + 0.144 +0.120

Estimación con ponderadores

Nivel de escolaridad	Número
NE1	0.6
NE2	0.4
Total	1

Gráfico 9.20d Número de personas con empleo remunerado en noviembre, por clasificación industrial

Estimación tradicional

Rama industrial	Número
DH	1
DJ	0
DK	0
Total	1

Estimación con ponderadores:
0.56 = 0.336 + 0.224

Estimación con ponderadores

Rama industrial	Número
DH	0.56
DJ	0.24
DK	0.20
Total	1

Gráfico 9.20e Número de personas con empleo remunerado en noviembre, por ocupación y nivel de escolaridad

Estimación tradicional

Nivel de escolaridad	Oc1	Oc2	Total
NE1	0	0	**0**
NE2	1	0	**1**
Total	**1**	**0**	1

Estimación con ponderadores:
0.48 = 0.336 + 0.144

Estimación con ponderadores

Nivel de escolaridad	Oc1	Oc2	Total
NE1	0.48	0.12	**0.60**
NE2	0.32	0.08	**0.40**
Total	**0.80**	**0.20**	1

Gráfico 9.20f Número de personas con empleo remunerado en noviembre, por nivel de escolaridad y rama industrial

Estimación tradicional

Rama industrial	NE1	NE2	Total
DH	**0**	1	**1**
DJ	0	0	**0**
DK	0	0	**0**
Total	**0**	**1**	1

Estimación con ponderadores:
0.336 está tomado directamente de la matriz de datos del gráfico 9.19f

Estimación con ponderadores

Rama industrial	NE1	NE2	Total
DH	0.336	0.224	**0.560**
DJ	0.144	0.096	**0.240**
DK	0.120	0.080	**0.200**
Total	**0.600**	**0.400**	1

Gráfico 9.20g Número de personas con empleo remunerado en noviembre, por ocupación y rama industrial

Estimación tradicional

Rama industrial	Oc1	Oc2	Total
DH	1	0	**1**
DJ	0	0	**0**
DK	0	0	**0**
Total	**1**	**0**	1

Estimación con ponderadores:
0.56 = 0.336 + 0.224

Estimación con ponderadores

Rama industrial	Oc1	Oc2	Total
DH	0.56	0.00	**0.56**
DJ	0.24	0.00	**0.24**
DK	0.00	0.20	**0.20**
Total	**0.80**	**0.20**	1

9.2.5 Consistencia entre diferentes variables

Para asegurar la consistencia cuando se usan distintas variables de valores múltiples en el sistema, se deben incluir los ponderadores correspondientes en el registro responsable de cada una de esas variables. Después se deben usar esos ponderadores en todo el sistema.

Algunos registros contienen muchas variables que deben cumplir ciertas condiciones de consistencia, como las entradas en un estado de pérdidas y ganancias. La consistencia se mantiene si se usan los mismos ponderadores para todas las variables, o si se recalculan primero las subentradas con diferentes ponderadores y luego se calculan los totales y las diferencias.

9.2.6 Consistencia entre estimaciones de diferentes registros

Cuando varios registros contienen la misma variable de valores múltiples, las estimaciones de los diferentes registros deben ser iguales. Para empezar, todas las poblaciones de registro deben ser consistentes. Así ocurre en el ejemplo siguiente: son las mismas personas en el registro de empleos y en el registro sobre personas. Asimismo, los establecimientos y las unidades empresariales son las mismas en el registro de empleos y en el registro de empresas.

En segundo lugar, todas las variables deben ser consistentes. En el ejemplo que veremos a continuación, las variables *rama industrial* y *jornada* (de trabajo) son consistentes en los diferentes registros. Todos los registros contienen las mismas definiciones de variables y errores de medición porque las variables se importaron del registro original a otros registros.

La tercera condición para la consistencia es que en las estimaciones se usen ponderadores, de modo que se considere toda la información de la variable de valores múltiples *rama industrial*. Si sólo se usa el valor más importante de *rama industrial* en vez de ponderadores, las estimaciones contendrán errores. *Estos errores de agregación diferirán de un registro a otro y, por consiguiente, las estimaciones serán inconsistentes.* Veamos esto en el siguiente ejemplo.

1. Integración de datos de cuatro registros: datos disponibles

El ejemplo consiste en un sistema de registros formado por cuatro registros: uno sobre cinco personas (NIP1–NIP5), uno de empleos con seis empleos (Em1–Em6), uno de establecimientos con tres establecimientos (Es11, Es21, Es22) y uno de empresas con dos unidades empresariales (UE1, UE2). En el gráfico 9.21 se muestra toda la información disponible antes de cualquier integración de datos.

Gráfico 9.21 Información disponible en cuatro registros antes de su integración

1. Registro sobre personas

Persona
NIP1
NIP2
NIP3
NIP4
NIP5

2. Registro de empleos

Empleo	Empresa	Establecimiento	Persona	$Jornada_E$
Em1	UE1	Es11	NIP1	0.3
Em2	UE1	Es11	NIP2	1.0
Em3	UE2	Es21	NIP3	1.0
Em4	UE2	Es21	NIP4	1.0
Em5	UE2	Es22	NIP5	1.0
Em6	UE2	Es22	NIP1	0.2

La persona NIP1 tiene dos empleos, Em1 y Em6, y trabaja 50% de un empleo de tiempo completo.

La información sobre las proporciones de las ramas industriales dentro de cada establecimiento se importará a los registros 1 y 2 cuando se creen los objetos combinados y el ponderador agregado se denominará w_{Ind} en el gráfico 9.23.

3. Registro de establecimientos

Establecimiento	Empresa	Rama industrial 1	%	Rama industrial 2	%
Es11	UE1	A	60	B	40
Es21	UE2	C	100		
Es22	UE2	D	100		

4. Registro de empresas

Empresa	Establecimiento 1	Establecimiento 2	Rama industrial 1	%	Rama industrial 2	%
UE1	Es11		A	60	B	40
UE2	Es21	Es22	C	62.5	D	37.5

2. Estimaciones tradicionales y errores de agregación

El gráfico 9.22 muestra el método tradicional de calcular el número de empleados de tiempo completo por rama industrial. En todos los registros, sólo se usa la *rama industrial más importante* para cada establecimiento, unidad empresarial, empleo o persona. De este modo, se fuerza a la *rama industrial* a volverse una variable de valor único: las estimaciones del número de empleados de tiempo completo por rama industrial será distinto a causa de los diferentes errores de agregación.

Gráfico 9.22 Estimación tradicional en un sistema de registros luego de su integración

1. Registro de personas empleadas

Persona	Rama industrial	Jor_P	ponderador
NIP1	A	0.5	1
NIP2	A	1.0	1
NIP3	C	1.0	1
NIP4	C	1.0	1
NIP5	D	1.0	1
Total		**4.5**	**5**

2. Registro de empleos

Empleo	Empresa	Establecimiento	Persona	Jor_E	Rama industrial	ponderador
Em1	UE1	Es11	NIP1	0.3	A	1
Em2	UE1	Es11	NIP2	1.0	A	1
Em3	UE2	Es21	NIP3	1.0	C	1
Em4	UE2	Es21	NIP4	1.0	C	1
Em5	UE2	Es22	NIP5	1.0	D	1
Em6	UE2	Es22	NIP1	0.2	D	1
Total				**4.5**		**6**

En todos los registros del gráfico 9.22, la variable *rama industrial* se refiere a la principal rama industrial. Cada registro también tiene un solo renglón por objeto (persona, empleo, establecimiento o unidad empresarial).

La variable Jor_E, *jornada de trabajo*, del registro de empleos se importa a todos los demás registros. Las variables derivadas Jor_P, *jornada para persona*, Jor_{Es}, *jornada para establecimiento*, *y* Jor_{UE}, *jornada para unidad empresarial*, se crean por agregación.

Con estos cuatro registros, se puede estimar el volumen de trabajo por rama industrial sumando las variables:

Jor_P , Jor_E , Jor_{Es} *y* Jor_{UE}

Como hay distintos errores de agregación, las tablas estimadas con volumen de trabajo por rama industrial serán diferentes si se usan diferentes registros.

Estas tablas se comparan en el gráfico 9.24.

3. Registro de establecimientos

Establecimiento	Rama industrial	Jor_{Es}	ponderador
Es11	A	1.3	1
Es21	C	2.0	1
Es22	D	1.2	1
Total		**4.5**	**3**

4. Registro de empresas

Empresa	Establecimiento	Rama industrial	Jor_{UE}	ponderador
UE1	Es11	A	1.3	1
UE2	Es21	C	3.2	1
Total			**4.5**	**2**

3. Estimación con objetos combinados

En todos los registros, se crean objetos combinados mediante *objeto · rama industrial* y se calculan los ponderadores para los objetos combinados. En el gráfico 9.23, las variables *rama industrial* y w_{Ind} se importaron del registro de establecimientos al registro de empleos. Sumando $Jor_E \cdot w_{Ind}$ en el registro de empleos, se obtienen la *jornada laboral para personas*, la *jornada laboral para establecimientos* y la *jornada laboral para unidades empresariales*.

Grafico 9.23 Estimaciones consistentes con ponderadores en un sistema de registros

1. Registro de Personas Empleadas

Persona	Rama industrial	Jor_P	w_{Ind}	w_E	$w_P = w_{Ind} \cdot w_E$	$Jor_P \cdot w_P$
NIP1	A	0.50	0.6	0.6	0.36	0.18
NIP1	B	0.50	0.4	0.6	0.24	0.12
NIP1	D	0.50	1	0.4	0.4	0.20
NIP2	A	1.00	0.6	1	0.6	0.60
NIP2	B	1.00	0.4	1	0.4	0.40
NIP3	C	1.00	1	1	1	1.00
NIP4	C	1.00	1	1	1	1.00
NIP5	D	1.00	1	1	1	1.00
Total					**5**	**4.5**

2. Registro de empleos

Empleo	Empresa	Establecimiento	Persona	Jor_E	Rama industrial	w_{Ind}	$Jor_E \cdot w_{Ind}$
Em1	UE1	Es11	**NIP1**	0.30	A	0.6	0.18
Em1	UE1	Es11	**NIP1**	0.30	B	0.4	0.12
Em2	UE1	Es11	NIP2	1.00	A	0.6	0.60
Em2	UE1	Es11	NIP2	1.00	B	0.4	0.40
Em3	UE2	Es21	NIP3	1.00	C	1	1.00
Em4	UE2	Es21	NIP4	1.00	C	1	1.00
Em5	UE2	Es22	NIP5	1.00	D	1	1.00
Em6	UE2	Es22	**NIP1**	0.20	D	1	0.20
Total						**6**	**4.5**

Comentarios sobre el registro de empleos:

El registro contiene datos sobre seis empleos correspondientes a 4.5 empleos de tiempo completo.

Los empleos Em1 y Em2 se dividen en dos objetos combinados cada uno, pues el Es11 está activo tanto en la rama industrial A como en la B. Los ponderadores 0.6 y 0.4 para estas dos ramas están tomados del registro de establecimientos. Jor_E se refiere a la jornada de trabajo de cada empleo.

Comentarios sobre el registro de empleados - personas:

Si se suma $Jor_E \cdot w_{Ind}$ para la persona NIP1 en el registro de empleados, se obtiene como resultado 0.18 + 0.12 + 0.20 = 0.50. Este valor se vuelve Jor_P para NIP1 en el registro de empleados.

Se forman tres objetos combinados correspondientes a tres ramas industriales para NIP1. Tanto la rama industrial como el empleo y el establecimiento son variables de valores múltiples para las personas.

Los ponderadores para objetos combinados se forman multiplicando w_{Ind} por w_E, donde w_{Ind} está tomado del registro de empleos y w_E se calcula como la participación de cada empleo de todos los empleos que tiene la persona. Para NIP1, el empleo Em1 tiene el ponderador (0.18+0.12)/(0.18+0.12+0.20) = 0.6

3. Registro de establecimientos

Establecimiento	Rama industrial	Jor_{Es}	w_{Ind}	$Jor_{Es} \cdot w_{Ind}$
Es11	A	1.3	0.6	0.78
Es11	B	1.3	0.4	0.52
Es21	C	2.0	1	2.00
Es22	D	1.2	1	1.20
Total			**3**	**4.5**

4. Registro de empresas

Empresa	Establecimiento	Rama industrial	Jor_{UE}	w_{Ind}	$Jor_{UE} \cdot w_{Ind}$
UE1	Es11	A	1.3	0.6	0.78
UE1	Es11	B	1.3	0.4	0.52
UE2	Es21	C	3.2	0.625	2.00
UE2	Es22	D	3.2	0.375	1.20
Total				**2**	**4.5**

4. Estimación consistente de empleados de tiempo completo por rama industrial

Los cuatro registros presentados en el gráfico 9.23 dan exactamente las mismas estimaciones sobre la distribución del volumen de trabajo, medido como número de empleados de tiempo completo, entre las diferentes ramas industriales. Esto se debe a que se utiliza toda

la información sobre la rama industrial cuando se han aplicado ponderadores en teoría consistentes en los cuatro registros.

En la tabla 1 del gráfico 9.24 se comparan diferentes estimaciones de empleados de tiempo completo por rama industrial. Las columnas "Registro de empleados", …, "Registro de empresas" son estimaciones calculadas de manera tradicional con los registros del gráfico 9.22. Las diferencias entre estas cuatro columnas sólo se deben a que el método de estimación es inadecuado. La última columna de la tabla 1 contiene las estimaciones basadas en objetos combinados que se llevaron a cabo a partir de los registros del gráfico 9.23. La tabla 2 del gráfico 9.24 muestra la estimación tradicional usando el registro sobre personas del gráfico 9.22 y, en la columna de la derecha, se presentan las estimaciones para las que se usaron los ponderadores w_P del registro sobre personas del gráfico 9.23. Por ejemplo, para la rama industrial A, la suma de w_P en el registro sobre personas es 0.36 + 0.6 = 0.96 en la tabla 2.

Gráfico 9.24 Comparación de los diferentes tipos de estimación

Tabla 1
Número de empleados de tiempo completo por rama industrial

Rama industrial	Estimación tradicional, gráfico 9.22				Estimación con objetos combinados, registros 1 a 4 del gráfico 9.23
	Registro de empleados	Registro de empleos	Registro de establecimientos	Registro de empresas	
A	1.5	1.3	1.3	1.3	0.78
B	0.0	0.0	0.0	0.0	0.52
C	2.0	2.0	2.0	3.2	2.00
D	1.0	1.2	1.2	0.0	1.20
Total	**4.5**	**4.5**	**4.5**	**4.5**	**4.5**

Tabla 2
Número de personas por rama industrial

Tradicional, gráfico 9.22	Con objetos combinados
Registro de personas empleadas	Registro de personas empleadas, gráfico 9.23
2	0.96
0	0.64
2	2.00
1	1.40
5	**5.00**

En consecuencia, el método de estimación tradicional arroja inconsistencias en las estadísticas basadas en registros que se obtienen. Por otro lado, las diferencias en las definiciones de la población y las variables pueden ocasionar otras inconsistencias en los registros reales.

Hay una cuarta razón por la que las estadísticas basadas en registros de diferentes productos pueden ser inconsistentes. La tabla anterior muestra los efectos de las diferencias relacionadas con el contenido. En la tabla 1 se describe a *empleados de tiempo completo* y en la tabla 2 a *personas*. En las estadísticas sobre personas y sobre el mercado laboral, es común que se describa a personas, pero en el caso de las estadísticas económicas es más frecuente medir volúmenes y empleados de tiempo completo.

9.2.7 Variables de valores múltiples: ¿qué se hace en la práctica?

Ejemplo: cambio de rama industrial

En la sección 9.2.2 veremos la variable de valores múltiples *clasificación industrial*. En la metodología tradicional, todas las actividades de una empresa con varias clasificaciones industriales se asignan a la principal clasificación industrial. Esto ocasiona cambios de nivel en las series de tiempo cuando cambia la principal clasificación industrial de una empresa. El grado de alteración de la serie de tiempo dependerá de la importancia de la empresa dentro de la rama industrial o la región que le corresponda.

El ejemplo siguiente se basa en un caso real, sólo adaptamos ligeramente los datos. En los años 1 a 3, la empresa ABC, S.A. ha llevado a cabo actividades en varias ramas industria-

les, pero alrededor de 60% de la facturación corresponde a la rama industrial R. En el año 4, ABC, S.A. compró una empresa con actividades en otra rama. El cambio de propiedad ocurrió a partir del cuarto trimestre del año 4.

La columna (2) del gráfico 9.25 contiene la facturación en millones de coronas suecas para todas las empresas de la rama industrial R, salvo ABC, S.A. Las columnas (3) y (4) contienen la facturación total para ABC, S.A. La columna (5) se tomó del Registro Empresarial y muestra la participación dentro de la rama industrial R.

Aplicando el método de estimación tradicional, toda la facturación de la empresa se asigna a la rama industrial R en los años 1 a 4. Obsérvese que la rama industrial R se mantiene como principal clasificación industrial durante todo el año 4 porque con esta metodología los cambios sólo se aplican al final del año. A partir del año 5, no se asigna ninguna facturación de la empresa a la rama industrial R. Sumando las columnas (2) y (7), obtenemos la serie de tiempo de la columna (8), que contiene un cambio de nivel.

Gráfico 9.25 Estimación de la facturación en una rama industrial aplicando dos métodos

Año Q	Rama industrial R excepto ABC, S.A.	ABC, S.A. antes de la adquisición	ABC, S.A. después de la adquisición	Participación de ABC, S.A. en la rama industrial R	Ponderador con estimación tradicional	Contribución de ABC, S.A. a la rama industrial R, estimación tradicional	Estimación tradicional de la rama industrial R	Contribución de ABC, S.A. a la rama industrial R con objetos combinados	Rama industrial R con objetos combinados
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1 1	7 684	7 354		0.60	1	7 354	15 038	4 412	12 096
1 2	7 086	7 086		0.60	1	7 086	14 172	4 252	11 338
1 3	8 142	6 788		0.60	1	6 788	14 930	4 073	12 215
1 4	9 853	8 387		0.60	1	8 387	18 240	5 032	14 885
...	...	...		...	...	...	...	...	...
4 1	13 071	9 259		0.57	1	9 259	22 330	5 278	18 349
4 2	13 127	9 509		0.57	1	9 509	22 636	5 420	18 547
4 3	11 253	9 499		0.57	1	9 499	20 752	5 414	16 668
4 4	12 921		15 881	0.21	1	15 881	28 802	3 335	16 256
5 1	12 782		12 397	0.21	0	000	12 782	2 603	15 385
5 2	13 360		12 634	0.21	0	000	13 360	2 653	16 013
5 3	11 098		11 621	0.21	0	000	11 098	2 440	13 538
5 4	12 888		13 209	0.21	0	000	12 888	2 774	15 662

De acuerdo con la metodología de estimación basada en objetos combinados, las participaciones de la columna (5) se deben usar como ponderadores al estimar la facturación para la empresa ABC, S.A. Multiplicando la facturación de la empresa en las columnas (3) y (4) por los ponderadores de la columna (5), obtenemos la parte de la facturación correspondiente a la rama industrial R. Al sumar esta parte de la facturación con la columna (9), que indica la facturación para las otras empresas de la rama industrial R, obtenemos la serie de tiempo de la columna (10), que describe la facturación de la rama industrial sin ningún cambio de nivel en la serie de tiempo.

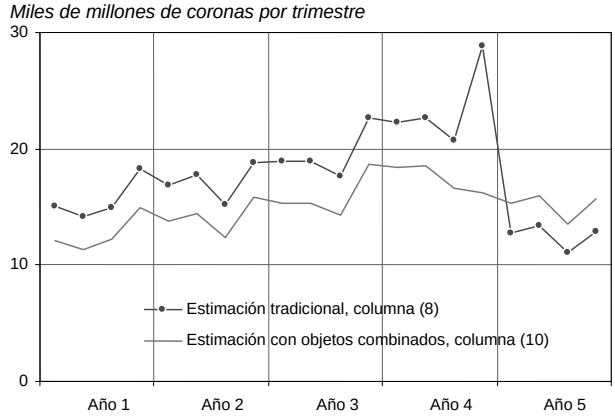
La serie de facturación estimada tradicionalmente se ve afectada por un cambio de nivel en la serie de tiempo causado por errores de agregación modificados abruptamente. Para el periodo año 1 a año 4, el método tradicional sobrestima la facturación de la rama industrial. Este error adquiere mayor relevancia en el cuarto trimestre del año 4. A partir del año 5, se subestima la facturación de esa rama.

La serie de la columna (10) se estimó usando los ponderadores de la columna (5). Si bien no son perfectos, estos ponderadores son considerablemente mejores que los ponderadores tradicionales de la columna (6).

Mediante el uso de ponderadores, los errores se pueden reducir bastante y la calidad de la serie de tiempo aumenta sustancialmente.

En el gráfico 9.26 se comparan los métodos de estimación.

Gráfico 9.26 Facturación en una rama industrial, dos estimaciones
Miles de millones de coronas por trimestre



Variables de valores múltiples: resumen de recomendaciones

Hay una serie de variables importantes en el sistema de registros que tienen valores múltiples. La manera en que se manejan actualmente estas variables en ocasiones puede producir estimaciones con errores de agregación. Estos errores se pueden reducir usando objetos combinados y ponderadores en las estimaciones. En esta sección, planteamos diferentes problemas de estimación que pueden presentarse con las variables de valores múltiples y recomendamos soluciones. Ya vimos en el ejemplo anterior sobre el cambio de rama industrial que métodos relativamente sencillos, como el uso de ponderadores, pueden mejorar la calidad aunque los ponderadores no sean del todo perfectos.

Otra ventaja importante del método de estimación que presentamos en esta sección (sección 9.2), considerando que las inconsistencias representan hoy un gran problema, estriba en que se puede lograr consistencia entre las estadísticas económicas para diferentes tipos de unidades empresariales.

9.2.8 Encuestas por muestreo con variables de registro de valores múltiples

Cuando en una encuesta por muestreo se usan variables de valores múltiples tomadas de registros, los ponderadores para el objeto combinado del registro se deben combinar con ponderadores de muestreo. A continuación, se deben calcular los totales con todos los objetos combinados pertenecientes a la unidad de muestreo correspondiente.

Ejemplo: En la Encuesta sobre Fuerza Laboral, el entrevistador pregunta sobre el establecimiento, cuya clasificación industrial se codifica usando la información del Registro Empresarial. Tradicionalmente, sólo se usa la rama industrial principal, lo que ocasiona errores de agregación en esa encuesta. Para evitar estos errores, en la matriz de datos de la encuesta por muestreo se debe usar el objeto combinado *establecimiento* • *rama industrial* del registro.

Los ponderadores d_i y g_i son los ponderadores de muestreo comunes y w_i es el ponderador para la variable de valores múltiples que describe la sección industrial en el Registro Empresarial.

$$\hat{Y} = \sum_{i=1}^{r_r} d_i g_i w_i y_i \qquad r_r \text{ es el número de objetos combinados que responden} \qquad (5)$$

Supongamos que se ha seleccionado a una persona, *NIP3*, que trabaja en el establecimiento Es4, el cual está activo en dos ramas industriales, DJ y DH, con una participación de 60% en DJ y 40% en DH. El ponderador de muestreo común $d_i\, g_i$ es igual a 353.8 para esta persona.

Gráfico 9.27 Matriz de datos para la Encuesta sobre Fuerza Laboral

Persona	Con empleo remunerado	Horas trabajadas	Establecimiento	Rama industrial	$d_i \cdot g_i$	w_{Ind}
Estimación tradicional						
NIP3	Sí	36	Es4	DJ	353.8	-
Estimación con objetos combinados						
NIP3	Sí	36	Es4	DJ	353.8	0.6
NIP3	Sí	36	Es4	DH	353.8	0.4

¿Cómo contribuyen a estas estimaciones los valores de la persona *NIP3*?

Con el *método de estimación tradicional*:

– el número de personas empleadas en la rama industrial DJ aumenta 353.8;
– el número de horas trabajadas en la rama industrial DJ aumenta 353.8 • 36 = 12 735.8 horas;
– la rama industrial DH no crece, ni en número de personas empleadas ni en número de horas trabajadas.

Con el *método de estimación presentado en la fórmula (5)*:

– el número de personas empleadas en la rama industrial DJ aumenta 353.8 • 0.6 = 212.3;
– el número de personas empleadas en la rama industrial DH aumenta 353.8 • 0.4 = 141.5;
– el número de horas trabajadas en la rama DJ aumenta 353.8 • 0.6 • 36 = 7 642.1;
– el número de horas trabajadas en la rama DH aumenta 353.8 • 0.4 • 36 = 5 094.7.

Las encuestas por muestreo que usan estimaciones específicas por rama industrial de acuerdo con la fórmula (5) presentan menos errores, incluidos errores de muestreo, que si usaran estimaciones tradicionales.

9.2.9 Combinación de datos de operativos de encuestas y administrativos en el Registro Empresarial

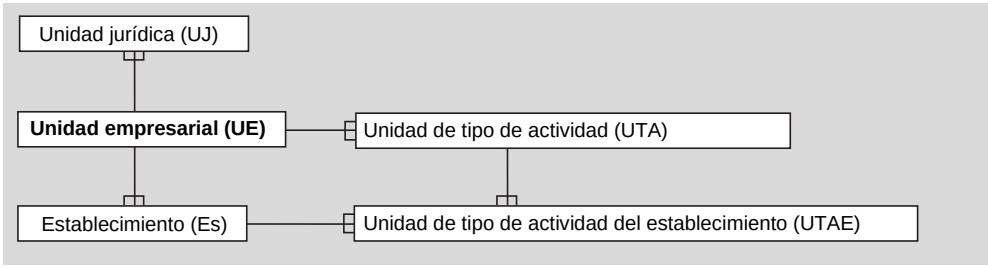
En la sección 2.2.3 describimos los diferentes tipos de objetos incluidos en el Registro Empresarial. Los datos administrativos relacionados con unidades jurídicas (UJ) y establecimientos (Es) se entregan a la Oficina de Estadística de Suecia.

Para crear unidades empresariales (UE) que tengan un significado estadístico, algunas unidades jurídicas (UJ) deben combinarse entre unidades empresariales, de modo que cada unidad empresarial conste de uno o más establecimientos.

Algunas unidades empresariales se dividen en unidades de tipo de actividad (UTA), que deben ser lo más *específicas posible respecto a la rama industrial*. Además, algunos establecimientos se dividen en unidades de tipo de actividad del establecimiento (UTAE) específicas de una rama industrial. Estas divisiones en unidades específicas de una rama industrial sólo se aplican actualmente cuando es *posible captar datos* sobre estas UTA y UTAE. En su trabajo con todas estas unidades, la Oficina de Estadística de Suecia acos-

tumbra comunicarse con las empresas, pero sólo entra en contacto con un número limitado de empresas grandes, por lo que es muy reducido el número de UE, UTA y UTAE creadas.

Gráfico 9.28 Tipos de objetos en el Registro Empresarial



En la sección 9.2, introdujimos el concepto *objeto combinado* para fines de estimación. Los métodos de estimación para la variable de valores múltiples *rama industrial* se basan en el hecho de que por cada combinación de rama industrial y unidad empresarial o establecimiento se crean objetos combinados. Éstos son objetos derivados que se crearon sin necesidad de cuestionarios enviados por correo o contactos telefónicos con las empresas; son totalmente específicos de una rama industrial, pero su propósito no es la captación de datos.

En algunas encuestas empresariales se usa una combinación de datos captados mediante cuestionarios en el caso de empresas grandes y datos administrativos para el resto de la población. En las empresas grandes se captan datos específicos por rama industrial sobre las unidades de tipo de actividad (UTA) o las unidades de tipo de actividad del establecimiento (UTAE).

Estos datos captados se pueden combinar con datos administrativos para los objetos combinados *unidad empresarial* • *rama industrial* o *establecimiento* • *rama industrial* con el propósito de generar estimaciones específicas para cada rama industrial.

Ejemplo: Encuesta de facturación en la que se combinan datos captados y datos administrativos

El registro empresarial de una región pequeña consta de cinco unidades jurídicas UJ1–5. Las UJ1–3 pertenecen al mismo grupo consolidado y se combinan en una unidad empresarial, UE1. Después de conversaciones entre representantes del Registro Empresarial y este grupo, se decidió dividir la unidad empresarial en dos unidades de tipo de actividad, UTA1 y UTA2, a las que se envían cuestionarios. Para el resto de la población se usan las unidades administrativas y los datos administrativos.

Gráfico 9.29 Encuesta de empresas con datos captados y datos administrativos

ID UJ	ID UE	ID UTA	Rama industrial 1	%	Rama industrial 2	%	Facturación	Fuente
UJ1–3	UE1	UTA1	DH	100			450	Datos captados
UJ1–3	UE1	UTA2	DJ	100			300	Datos captados
UJ4	UE2		DH	60	DJ	40	250	Datos administrativos
UJ5	UE3		DJ	70	DH	30	150	Datos administrativos

Las estimaciones tradicionales de facturación por rama industrial son:

– facturación DH = 450 + 250 = 700

– facturación DJ = 300 + 150 = 450

Para reducir los errores de agregación, es mejor crear unidades combinadas correspondientes a cada combinación *unidad empresarial* • *rama industrial* para la UJ4 (UE2) y la UJ5 (UE3).

Gráfico 9.30 Encuesta de empresas con datos colectados y datos administrativos

ID UJ	ID UE	ID UTA	Rama industrial	ponderador	Facturación	Fuente
UJ1–3	UE1	UTA1	DH	1	450	Datos captados
UJ1–3	UE1	UTA2	DJ	1	300	Datos captados
UJ4	UE2	UTA3	DH	0.6	250	Datos administrativos
UJ4	UE2	UTA4	DJ	0.4	250	Datos administrativos
UJ5	UE3	UTA5	DJ	0.7	150	Datos administrativos
UJ5	UE3	UTA6	DH	0.3	150	Datos administrativos
Total				4		

Usando los objetos combinados, las estimaciones de facturación por rama industrial son:

– facturación DH = 450 + 0.6 • 250 + 0.3 • 150 = 645

– facturación DJ = 300 + 0.4 • 250 + 0.7 • 150 = 505

Se crean dos objetos combinados para UJ4 y UJ5, respectivamente. Estos cuatro objetos combinados son cuatro unidades de tipo de actividad (UTA) derivadas. Las estimaciones por rama industrial con menores errores de agregación se calculan mediante los ponderadores del gráfico 9.30, donde los datos captados siempre tendrán el ponderador 1.

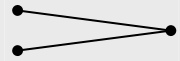
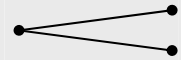
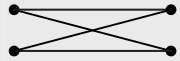
9.3 VINCULACIÓN DE LAS SERIES DE TIEMPO MEDIANTE OBJETOS COMBINADOS

En este capítulo hemos visto métodos de estimación para variables de valores múltiples. Cuando un objeto tiene varios valores para una variable de valores múltiples, se crea un objeto combinado para cada combinación de un objeto y un valor posible. Cada uno de estos objetos combinados tendrá un ponderador que se usará cuando se estimen los valores en las celdas de una tabla. Se puede usar el mismo método para los cambios de nivel de la series de tiempo ocasionados por *clasificaciones modificadas*, en las que no son completamente comparables las categorías nuevas y las anteriores. Una de las ventajas del método que presentamos aquí, basado en la creación de objetos combinados, radica en que no es necesario vincular primero las series en el nivel macro.

Cuando, por ejemplo, se modifica el sistema de clasificación industrial, es posible que un código antiguo se remplace por uno o varios códigos nuevos. Por lo general, los códigos de versiones anteriores del Registro Empresarial pueden remplazarse por los códigos nuevos si se cuenta con información sobre la gama de productos de las empresas en el registro antiguo. De este modo, la gama de productos se puede convertir a los códigos industriales nuevos. Bayard y Klimek (2003) han aplicado este método, en combinación con otros, en la vinculación de series de tiempo en el nivel micro. Sin embargo, el método presentado en esta sección requiere menos datos y puede usarse en caso de que falte información sobre la gama de productos.

Cuando se modifica un sistema de clasificación, los códigos anteriores y los nuevos pueden tener las relaciones que se muestran en el gráfico 9.31.

Gráfico 9.31 Conversión de los códigos anteriores a los nuevos

Relación entre los códigos anteriores y los nuevos	Código anterior	Clave de código	Código nuevo	Comentarios
Uno a uno	1		A	No hay problemas, el código 1 anterior se recodifica como nuevo código A.
Muchos a uno	2 3		B	No hay problemas, los códigos 2 y 3 anteriores se combinan en el nuevo código B.
Uno a muchos	4		C D	Hay problemas, ¿cómo se debe dividir el código 4 anterior en los códigos nuevos C y D?
Muchos a muchos	5 6		E F	Mismo problema anterior, ¿cómo se debe… – dividir el código 5 anterior en E y F? – dividir el código 6 anterior en E y F?

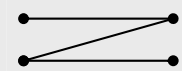
Una condición para el método que describimos en esta sección es que haya un registro para un momento o periodo particulares en el que cada objeto esté codificado de acuerdo con la clasificación anterior y la nueva.

Si la clasificación es una característica estable, será posible recodificar usando una nueva clasificación en el caso de objetos idénticos en las versiones anteriores del registro. Pero este tipo de reclasificación sólo es aplicable a la parte de la población de registro que se puede rastrear hacia atrás en el tiempo. Cuando *no* es posible reclasificar objeto por objeto de esta forma, se puede recurrir al método de objetos combinados que describimos más adelante.

Los problemas surgen cuando los objetos con códigos anteriores se reclasifican con uno de varios códigos nuevos. Esto se resuelve combinando cada uno de los objetos con cada código nuevo posible; posteriormente, a estos objetos combinados se les asignan los ponderadores que se usan cuando se estiman los valores en las celdas de la tabla.

Ejemplo: Código de clasificación industrial modificado en el Registro Empresarial
En este ejemplo, utilizamos datos ficticios. La *clasificación industrial* se modifica en el año 3; los códigos 1 y 2 anteriores se remplazan con los nuevos códigos A y B. Para el año 3, a todas las unidades empresariales (UE) se les asignan tanto los códigos anteriores como los nuevos. Para los años 1 y 2, sólo existen los códigos industriales anteriores. En el gráfico 9.32 se muestra la correspondencia entre los códigos anteriores y los nuevos.

Gráfico 9.32 Clave entre los códigos de clasificación industrial anteriores y los nuevos

Relación entre códigos	Código anterior	Clave de código	Código nuevo	Comentarios
Muchos a muchos	1 2		A B	El 1 anterior se convierte en el nuevo A, pero el 2 anterior se debe dividir entre el nuevo A y el nuevo B.

El código 1 anterior no presenta problemas, pues siempre se debe codificar como A en la clasificación nueva. Pero el código 2 anterior sí plantea dificultades: ¿a la empresa que tenía un código 2 anterior se le debe asignar el código nuevo A o B? Con el método de objetos combinados, dos objetos combinados con los códigos A y B remplazan a las empresas con el código 2, respectivamente.

En el gráfico siguiente vemos el Registro Empresarial para tres años con datos sobre el número de empleados (*Núm. empleados*) y el *valor agregado* correspondiente a cada unidad empresarial. UE8 y UE13 tienen dos clasificaciones industriales y, por consiguiente, dos objetos combinados cuyos ponderadores se conocen.

Gráfico 9.33

Registro empresarial, año 1

Empresa	Rama industrial	Núm. empleados	Valor agregado	w_i
UE1	2	10	15	1
UE2	1	20	50	1
UE3	2	50	100	1
UE4	1	30	75	1
UE5	2	10	15	1
UE6	1	20	40	1
UE7	2	20	40	1
UE8	1	60	105	0.6
UE8	2	60	105	0.4
UE9	1	40	80	1
UE10	2	20	40	1
UE11	1	10	25	1
UE12	2	10	15	1

Registro empresarial, año 2

Empresa	Rama industrial	Núm. empleados	Valor agregado	w_i
UE1	2	10	15	1
UE2	1	20	50	1
UE3	2	50	100	1
UE13	1	40	80	0.5
UE13	2	40	80	0.5
UE14	1	10	25	1
UE15	2	20	40	1
UE16	1	40	80	1
UE17	2	30	60	1
UE18	1	40	80	1
UE19	2	20	40	1
UE20	1	10	25	1
UE21	2	10	20	1

Registro empresarial, año 3

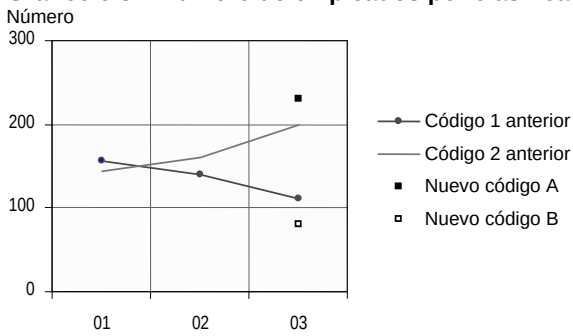
Empresa	Rama industrial anterior	Rama industrial nuevo	Núm. empleados	Valor agregado	w_i
UE1	2	B	10	15	1
UE2	1	A	20	50	1
UE3	2	A	50	100	1
UE13	1	A	40	80	0.5
UE13	2	B	40	80	0.5
UE14	1	A	10	25	1
UE15	2	A	20	40	1
UE22	1	A	10	25	1
UE23	2	B	40	60	1
UE24	1	A	30	75	1
UE25	2	A	50	100	1
UE26	1	A	20	50	1
UE27	2	B	10	15	1

Con la información de estos tres registros, es posible formar series de tiempo de acuerdo con la antigua clasificación industrial.

Los valores para el año 3 también se pueden calcular de acuerdo con la clasificación industrial nueva. Como las UE8 y UE13 realizan actividades dentro de dos clasificaciones industriales, se forma el objeto combinado *UE • rama industrial* y se obtienen los ponderadores w_i, que se usan en las estimaciones para estas series de tiempo, según se describió en la sección 9.2.

En el próximo gráfico, vemos el cambio de nivel de la serie de tiempo para el año 3, ¿pero cómo se pueden calcular los valores corregidos para años anteriores?

Gráfico 9.34 Número de empleados por clasificación industrial, años 1–3



En el gráfico siguiente vemos cómo hacer la vinculación en este caso. Para el año 3, conocemos tanto el código industrial anterior como el nuevo para todas las unidades empresariales del registro. Como la clasificación industrial es una variable relativamente estable,

usamos el nuevo código industrial del año 3 para los años 1 y 2. De esta forma, las UE1–3 reciben el nuevo código industrial para los años 1 y 2, mientras que las UE13–15 reciben los nuevos códigos para el año 2. ¿Qué se debe hacer con las empresas que no están incluidas en el registro del año 3? Las empresas con el código industrial 1 anterior no representan ningún problema y se les asigna el nuevo código A. Enseguida ilustramos estos primeros pasos del proceso de vinculación, que son muy sencillos.

Gráfico 9.35 Vinculación en el nivel micro: los pasos sencillos

Registro empresarial, año 1

Empresa	Rama industrial Ant.	Rama industrial Nvo.	Núm. empleados	Valor agregado	w_i ant.	w_i empleados	w_i valor agregado
UE1	2	B	10	15	1	1	1
UE2	1	A	20	50	1	1	1
UE3	2	A	50	100	1	1	1
UE4	1	A	30	75	1	1	1
UE6	1	A	20	40	1	1	1
UE9	1	A	40	80	1	1	1
UE11	1	A	10	25	1	1	1

Registro empresarial, año 2

Empresa	Rama industrial Ant.	Rama industrial Nvo.	Núm. empleados	Valor agregado	w_i ant.	w_i empleados	w_i valor agregado
UE1	2	B	10	15	1	1	1
UE2	1	A	20	50	1	1	1
UE3	2	A	50	100	1	1	1
UE13	1	A	40	80	0.5	0.5	0.5
UE13	2	B	40	80	0.5	0.5	0.5
UE14	1	A	10	25	1	1	1
UE15	2	A	20	40	1	1	1
UE16	1	A	40	80	1	1	1
UE18	1	A	40	80	1	1	1
UE20	1	A	10	25	1	1	1

Registro empresarial, año 3

Empresa	Rama industrial Ant.	Rama industrial Nvo.	Núm. empleados	Valor agregado	w_i
UE1	***2***	***B***	***10***	***15***	1
UE2	1	A	20	50	1
UE3	**2**	**A**	**50**	**100**	1
UE13	1	A	40	80	0.5
UE13	***2***	***B***	***40***	***80***	***0.5***
UE14	1	A	10	25	1
UE15	**2**	**A**	**20**	**40**	1
UE22	1	A	10	25	1
UE23	***2***	***B***	***40***	***60***	1
UE24	1	A	30	75	1
UE25	**2**	**A**	**50**	**100**	1
UE26	1	A	20	50	1
UE27	***2***	***B***	***10***	***15***	1

¿Qué se debe hacer con las empresas que no están en el registro del año 3 y tienen el código 2 anterior? Como el código 2 anterior se puede convertir en los nuevos códigos A y B, se crean objetos combinados para las UE5, 7, 8, 10, 12, 17, 19 y 21.

¿Cómo debemos generar ponderadores para estos objetos combinados? Estos ponderadores deben mostrar cómo se divide el código 2 anterior en los nuevos códigos A y B. Para el año 3 podemos usar los datos disponibles en el gráfico anterior a fin de calcular los ponderadores. Se generan diferentes ponderadores para las variables *número de empleados* y *valor agregado* porque las nuevas ramas industriales pueden tener una estructura distinta. Los cálculos se muestran en el gráfico 9.36. Se usaron los valores en negritas del gráfico anterior.

Gráfico 9.36 Generación de ponderadores para los objetos combinados

Rama industrial, código anterior	Rama industrial, código nuevo	Número de empleados	Valor agregado
2	A	50 + 20 + 50 = 120	100 + 40 + 100 = 240
2	B	10 + 40 • 0.5 + 40 +10 = 80	15 + 80 • 0.5 + 60 + 15 = 130
Total		200	370

– El ponderador w_i *anterior* produce estimaciones para la rama industrial anterior, mientras que w_i *empleados* y w_i *valor agregado* arrojan estimaciones para el número de empleados y el valor agregado de acuerdo con la clasificación industrial nueva.

– El número de empleados es 200, de los cuales las empresas dentro de la rama industrial A tienen 120, entonces 120/200 = 0.6.

Por lo tanto, w_i *empleados* es 0.6 para la rama industrial A y 0.4 para la B.

– El valor agregado es 370, del que 240 corresponde a las empresas dentro de la rama industrial A, entonces 240/370 = 0.65.

Por lo tanto, w_i *valor agregado* es 0.65 para la rama industrial A y 0.35 para la B.

Suponemos que estos ponderadores se pueden usar para los años 1 y 2, y mostramos los objetos combinados con ponderadores en el gráfico siguiente.

Gráfico 9.37 Vinculación en el nivel micro: uso de objetos combinados

Registro Empresarial, año 1

Empresa	Rama industrial		Núm. empleados	Valor agregado	w_i anterior	w_i empleados	w_i valor agregado
	Anterior	Nuevo					
UE5	2	A	10	15	1	0.6	0.65
UE5	2	B	10	15	0	0.4	0.35
UE7	2	A	20	40	1	0.6	0.65
UE7	2	B	20	40	0	0.4	0.35
UE8	1	A	60	105	0.6	0.6	0.6
UE8	2	A	60	105	0.4	0.24	0.26
UE8	2	B	60	105	0	0.16	0.14
UE10	2	A	20	40	1	0.6	0.65
UE10	2	B	20	40	0	0.4	0.35
UE12	2	A	10	15	1	0.6	0.65
UE12	2	B	10	15	0	0.4	0.35

Registro Empresarial, año 2

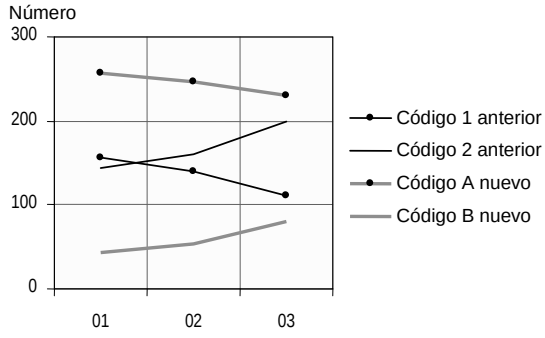
Empresa	Rama industrial		Núm. empleados	Valor agregado	w_i anterior	w_i empleados	w_i valor agregado
	Anterior	Nuevo					
UE17	2	A	30	60	1	0.6	0.65
UE17	2	B	30	60	0	0.4	0.35
UE19	2	A	20	40	1	0.6	0.65
UE19	2	B	20	40	0	0.4	0.35
UE21	2	A	10	20	1	0.6	0.65
UE21	2	B	10	20	0	0.4	0.35

Nota: Los ponderadores para UE8 con rama industrial 2 se calculan de la siguiente manera (celdas en blanco):

Nuevo A: 0.4 • 0.6 = 0.24 y 0.4 • 0.65 = 0.26
Nuevo B: 0.4 • 0.4 = 0.16 y 0.4 • 0.35 = 0.14

Combinando los datos de los gráficos 9.35 y 9.37 en un registro completo, se pueden calcular los valores de las series de tiempo vinculadas para *número de empleados* y *valor agregado* correspondientes a los años 1 y 2. En el siguiente gráfico, se comparan las series de tiempo relacionadas con la clasificación anterior y la nueva.

Gráfico 9.38 Empleados por clasificación industrial, años 1–3



En este ejemplo, los ponderadores usados para los cálculos se mantienen constantes en todos los años.

Si el vínculo se refiere a una rama industrial nueva establecida en años recientes, los ponderadores pueden disminuir hacia atrás en el tiempo.

CAPÍTULO 10

Calidad de las estadísticas basadas en registros

En Statistics Denmark (1995, capítulo 4) se hace un resumen de los requisitos que debe cumplir un sistema de registros para el buen aprovechamiento de los datos administrativos para propósitos estadísticos. Se señala que las estadísticas basadas en registros serán de alta calidad si: *el sistema de registros tiene una buena cobertura y contiene una amplia gama de tipos de objeto importantes, relaciones entre objetos y muchas variables.*

En la sección 1.6 mencionamos que hay tanto prejuicios como críticas legítimas respecto a las estadísticas basadas en datos administrativos. Quienes albergan prejuicios consideran que este tipo de estadísticas son de bajo costo, pero de mala calidad en comparación con las "verdaderas" estadísticas basadas en encuestas. ¿Qué motiva estas actitudes? Los medios de comunicación, como vemos en el ejemplo siguiente tomado de un importante diario sueco, muchas veces pueden propiciar esas actitudes.

Dagens Nyheter, marzo 2 de 2003

Mienten estadísticas sobre homicidios

Un nuevo sistema de cómputo puede estar mintiendo sobre las cifras que muestran el triple de casos de los que realmente han ocurrido.

Mikael Rying escribió un artículo sobre las diferencias entre el número de denuncias y homicidios registrados estadísticamente en Suecia y el número real de asesinatos.

En el periodo 1990–1998, en promedio sólo 57% de los casos denunciados fueron casos reales de asesinatos y homicidios culposos, señala Rying. Los otros casos podrían ser desde suicidio e intoxicación alcohólica hasta simples errores de codificación en las denuncias ante las autoridades.

Año	Mortalidad por causa de defunción: homicidio doloso y culposo	Delitos denunciados: homicidio doloso y culposo
96	110	199
97	94	157
98	98	185
99	108	188
00	90	175
Promedio para 1996–2000:	100	181
	100/181 = 55%	

Fuente: Anuario estadístico de Suecia, 2002–2003

Este ejemplo nos lleva a concluir que no es una buena idea usar sólo *una* fuente administrativa. El principio fundamental en el que se basa un sistema de registros es que se deben usar muchas fuentes para lograr un alto nivel de calidad y consistencia. Además, los datos administrativos no se deben usar directamente, sino que se deben validar y procesar para adecuarlos a las necesidades del producto estadístico de que se trate. En este caso, el problema con los delitos denunciados es que en la denuncia no se diferencia claramente entre distintos tipos de objetos: víctimas de homicidio, presuntos criminales y el homicidio en sí. Esto puede propiciar errores de codificación.

Otra fuente administrativa que podría usarse es el Registro de Causas de Defunción, que forma parte del sistema de registros de la Oficina de Estadística de Suecia y se coteja con el Registro de Población. El Registro de Causas de Defunción no está afectado por las fallas de calidad presentes en los datos sobre delitos denunciados. En comparación, es improbable

que una encuesta por muestreo produzca estadísticas sobre homicidios de mejor calidad —un cuestionario para todas las comisarías del país, o una muestra de ellas, no haría sino duplicar el suministro de datos en gran medida incorrectos—. Las dos encuestas basadas en registros mencionadas en este ejemplo deben tener información sobre su calidad, que pueda ser revisada de manera sencilla por los usuarios de la información, para evitar las interpretaciones incorrectas de los datos.

Otra opinión común sobre los datos administrativos es que los entrevistados sólo dan los datos que les benefician o ayudan a sus propósitos. En el ejemplo siguiente, es un hecho que a la mayoría de la gente le gustaría pagar la menor cantidad posible de impuestos, de modo que puede ocurrir que las deducciones sean mayores de lo que es justificable (del diario sueco Dagens Nyheter, julio 16 de 2003).

80 por ciento de las deducciones impositivas de los suecos son evasión fiscal

Los contribuyentes presentan sus declaraciones de impuestos con errores que ascienden a miles de millones. Reglas complicadas y una legislación poco clara han dificultado a las autoridades hacendarias del país verificar todas las deducciones. Los errores se encuentran sobre todo en las deducciones por concepto de operaciones bursátiles, honorarios de gestión y otros cargos relacionados con la compra venta de acciones.

Deducciones por compra venta de acciones

– Un tercio de las compras o ventas de acciones presentan errores.
– 700 000 contribuyentes declaran utilidades por alrededor de 50 mil millones de coronas y pérdidas por alrededor de 10 mil millones.
– Los errores en las declaraciones fiscales son difíciles de evaluar y ascienden a miles de millones de coronas.
– Muchos errores pasan inadvertidos a causa de la complejidad de las reglas.

Deducción por honorarios de gestión

– 125 000 contribuyentes dedujeron un total de 515 millones de coronas.
– 66% de estas deducciones contienen datos incorrectos.
– En total, los errores en las declaraciones fiscales pueden ascender a 90 millones de coronas.
– El error más común es el de la deducción por concepto de honorarios de gestores de fondos, pues los honorarios se deducen automáticamente.

Deducciones por otros gastos

– 700 000 contribuyentes dedujeron un total de 2 800 millones de coronas.
– 82% de estas deducciones contienen datos incorrectos.
– En total, los errores en las declaraciones fiscales pueden ascender a 700 millones de coronas.

El encabezado es tendencioso en varios sentidos: “ochenta por ciento” es una exageración y la “evasión fiscal” muchas veces se debe a interpretaciones incorrectas producto de la complejidad de las reglas:

– Deducciones por compra o venta de acciones: los errores son en buena medida involuntarios y se deben a la complejidad de las reglas.
– Deducciones por honorarios de gestión: los errores promedian 17% (90/515) y el más común de ellos puede ser involuntario, a causa de alguna mala interpretación.
– Deducciones por otros gastos: 82% de estas deducciones presentan errores, pero en promedio las deducciones son 25% (700/2 800) incorrectas.

Podemos tener otra perspectiva de estos errores si los comparamos con los ingresos totales de todos los contribuyentes, entonces el error es de 0.3 por ciento.

El hecho de que se hagan demasiadas deducciones en las declaraciones y de que, en consecuencia, los impuestos por pagar sean muy bajos no significa que las estadísticas del Registro de Ingresos sean de mala calidad, a pesar de que se basan en esas declaraciones. Supongamos que tenemos datos sobre una persona con deducciones demasiado altas en su declaración de impuestos, pero por lo demás su declaración es correcta:

Ingresos por empleo	257 600	Los ingresos son correctos.
Deducciones por otros gastos	25 500	La deducción es demasiado alta, pero aceptable.
Ingresos gravables	232 100	El ingreso gravable es incorrecto de acuerdo con las reglas fiscales, pero no se usa para fines estadísticos.
Impuestos	100 000	Los impuestos son incorrectos y demasiado bajos de acuerdo con las reglas fiscales, pero la cifra es correcta en términos estadísticos, porque son los impuestos que realmente pagó la persona.
Ingreso disponible	157 600	Correcto en términos estadísticos.

Basándonos en la información de esta persona, las cifras de la Oficina de Estadística de Suecia en relación con los ingresos percibidos no son incorrectas; tampoco son incorrectas las estadísticas relativas al ingreso disponible, pues son el resultado de calcular la diferencia entre los ingresos y los impuestos pagados realmente.

10.1 PROBLEMAS DE CALIDAD ESPECÍFICOS DE LAS ESTADÍSTICAS BASADAS EN REGISTROS

¿Qué es calidad estadística cuando hablamos de estadísticas basadas en registros? Platek y Särndal (2001) afirman que los estadísticos se han demorado en formular una teoría sobre la evaluación de la exactitud para estadísticas basadas en registros y que es necesaria esta teoría.

Holt (2001) señala que hay diferencias importantes entre las encuestas con una captación de datos propia y las estadísticas basadas en fuentes administrativas. Sostiene que en el caso de las estadísticas basadas en registros el aspecto más importante de la calidad no es la exactitud, sino la relevancia.

Por su parte, Nanopoulos (2001) opina que países como Dinamarca, que cuentan con un sistema de registros bien integrado, necesitan un aparato conceptual para los errores en las estadísticas, el cual será distinto del que requieran países en los que la mayoría de los datos estadísticos provienen de encuestas por muestreo y censos.

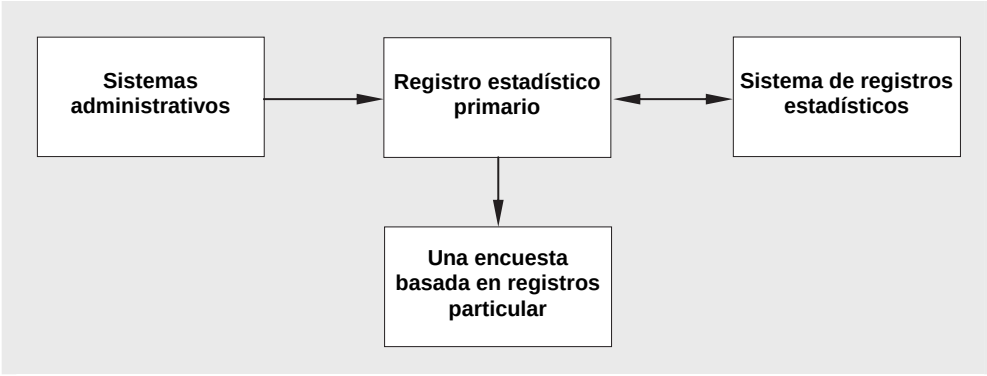
Nosotros concluimos que cuando se habla de la calidad de las estadísticas basadas en registros es importante considerar los siguientes factores:

– Es necesario distinguir entre encuestas con una captación de datos propia y las que se basan en registros. De lo contrario, se corre el riesgo de usar indiscriminadamente modelos de error tradicionales diseñados para encuestas por muestreo y censos.
– También se debe distinguir entre la calidad de una encuesta basada en registros y la de un registro estadístico, pues a este último se le pueden dar varios usos.

Las encuestas por muestreo y los censos se llevan a cabo teniendo en mente *un* uso en particular y, por lo general, los problemas de calidad tienen que ver con las estimaciones realizadas. En el caso de un registro estadístico, es posible que se le den *muchos* usos distintos, pues puede servir no sólo para encuestas actuales sino también para otras que se realicen en el futuro.

En el gráfico siguiente vemos la diferencia entre registro y encuesta basada en registros, así como dos elementos cruciales en relación con la calidad del registro: los sistemas administrativos y el sistema de registros estadísticos, las dos fuentes en las que se basa un registro estadístico.

Gráfico 10.1 Calidad en un registro y en una encuesta basada en registros



A semejanza de otras encuestas, la calidad de una encuesta basada en registros también se relaciona con *un* uso específico del registro y se centra en la calidad de las estimaciones, en particular su relevancia y exactitud en relación con el propósito de la encuesta. En este caso, describir la calidad se refiere a indicar si la calidad de la encuesta es *buena* o *mala*.

Sin embargo, la calidad del registro estadístico en sí no se relaciona con un uso en particular y, cuando se describe la calidad a este respecto, es importante especificar *qué características* tiene el registro, lo que permite saber los usos que podrían dársele. La calidad del registro afectará la calidad de las encuestas basadas en él y está determinada por tres factores:

– los sistemas administrativos en los que se basa el registro;
– las posibilidades que ofrece el sistema de registros estadísticos en cuanto a mejorar la cobertura, el contenido de las variables y la consistencia;
– el procesamiento realizado para producir el registro.

El primer factor que determina la calidad del registro son los *sistemas administrativos* en los que se basa (véase gráfico 10.1). Estos sistemas suelen ser únicos —la captación de los datos administrativos en el caso de un registro de población, por ejemplo, es totalmente distinta de aquella para los estados de resultados de las empresas—. La captación de datos en los sistemas administrativos habitualmente es distinto del que hace una oficina de estadística. Aunque en los sistemas administrativos participan informantes que llenan formularios, la presentación de información tiene condiciones propias y está sujeta a normas y reglamentos administrativos.

El segundo factor que determina la calidad del registro se relaciona con las posibilidades que ofrece el *sistema de registros*. El registro debe estar coordinado con el resto del sistema y el sistema en su conjunto debe funcionar con eficiencia.

El tercer factor determinante de la calidad del registro es el *procesamiento* efectuado cuando se creó el registro. ¿Cómo se definió la población de registro? ¿Cómo se determinó el contenido de las variables? ¿Cómo se validaron los datos?

Hasta ahora se ha considerado que el sistema administrativo y el procesamiento interno de los datos son específicos de cada registro, lo que ha dificultado el intercambio de experiencias y la formulación de una metodología; sin embargo, la calidad de un registro en particular afecta otras encuestas basadas en registros que usan ese registro. De este modo, por ejemplo, los valores perdidos para la variable *clasificación industrial* en el Registro Empresarial representa un problema para todos los registros y encuestas que incluyen esta

variable. Además, todos dependerán de que los responsables del Registro Empresarial solucionen este problema, pues de lo contrario los distintos registros no serán consistentes.

10.2 ERRORES EN LAS ENCUESTAS POR MUESTREO Y LAS ENCUESTAS BASADAS EN REGISTROS

La calidad de una encuesta por muestreo o un censo está determinada sobre todo por la eficiencia del proceso de captación de los datos. Este proceso se asemeja bastante en todas las encuestas con una captación de datos propia. Por ello, la siguiente lista de principales factores de calidad deberá aplicarse en mayor o menor grado a todas las encuestas por muestreo o censos:

– errores de muestreo (en el caso de encuestas por muestreo)
– no respuesta, y
– errores de medición.

El hecho de que se encuentren los mismos problemas metodológicos en todas las encuestas facilita el intercambio con los colegas, la formulación de metodología y la definición de lineamientos. Por otro lado, cabe señalar que las distintas encuestas por muestreo no repercuten unas en otras: la no respuesta en la Encuesta sobre Fuerza Laboral, por ejemplo, no afecta, por decir algo, a la Encuesta sobre Condiciones de Vida o la Encuesta sobre Entregas y Nuevos Pedidos en la Industria.

Aunque los factores de calidad más importantes probablemente no sean los mismos para todas las encuestas basadas en registros, en general la calidad de un registro afectará la calidad de otros. Por consiguiente, podemos establecer la siguiente comparación entre las encuestas basadas en la captación de datos y las basadas en registros:

Encuestas basadas en captación de datos:	**Encuestas basadas en registros:**
– Mismos problemas de calidad en todas las encuestas.	– Distintos problemas de calidad en las distintas encuestas.
– La calidad de una encuesta no afecta a otras encuestas.	– La calidad de una encuesta afecta a muchas otras encuestas.

Hasta el momento, no ha habido mucho intercambio de experiencias sobre la metodología de registros estadísticos y los problemas de calidad relacionados. No obstante, esperamos que una terminología y una perspectiva comunes estimulen ese intercambio.

10.2.1 Errores de muestreo y errores de integración

Durante mucho tiempo, se ha considerado que el error de muestreo es el más importante en las encuestas por muestreo. Por lo tanto, se han elaborado diseños muestrales y métodos de estimación para reducir este tipo de error. Cochran (1963) dedica doce de los trece capítulos de su libro a estos problemas; en el último capítulo, menciona los errores de medición y la no respuesta.

En las encuestas basadas en registros no hay una fase de muestreo, sino que predomina la *fase de integración*, en la que se integran datos de las distintas fuentes en un nuevo registro estadístico. En esta fase se crean la población de registro y los objetos derivados, se importan variables de distintas fuentes y se generan variables derivadas. Los tipos de errores originados en esta fase deben recibir el nombre de *errores de integración*, categoría que incluye errores de cobertura, errores de emparejamiento, valores perdidos por no empare-

jamiento (véase sección 8.1) y errores de agregación (véanse secciones 9.1 y 9.2). En la sección 10.4, presentamos 40 indicadores de calidad para las encuestas basadas en registros, de los cuales 22 corresponden a la fase de integración.

10.2.2 Errores en tabulados detallados

En el caso de las encuestas por muestreo, el error de muestreo determina qué tan detallado puede ser un tabulado para su publicación. Sin embargo, para las encuestas basadas en registros no hay un error de muestreo que nos indique si el tabulado es demasiado detallado para ser publicado. En las encuestas basadas en registros hay otros tipos de errores que son pequeños en un nivel más agregado, pero que ocasionan distorsiones si se publican tabulados demasiado detallados.

Clasificaciones jerárquicas

Los estándares y clasificaciones estadísticas importantes como las de ocupación, nivel de escolaridad y rama industrial, contienen cientos o miles de categorías en el nivel más detallado; no obstante, en muchos casos la calidad no es buena en ese nivel. En las encuestas por muestreo, el error de muestreo sólo permite que se publiquen estimaciones en los niveles más agregados. En el caso de las encuestas basadas en registros y los censos, se debe echar mano de otros métodos para evitar el uso o la publicación de tabulados de baja calidad.

Mencionamos la solución a este problema en la sección 2.5, donde introdujimos el concepto *variable estandarizada*. Las clasificaciones importantes se deben definir como variables estandarizadas, cuya documentación y calidad es responsabilidad de los registros donde se crearon por primera vez dichas variables dentro del sistema de registros.

Por ejemplo, los responsables del Registro Empresarial deben examinar la calidad de la variable *rama industrial*. También deben agrupar las categorías detalladas de esta variable y crear una variable derivada, que sea adecuada para las tablas basadas en registros. Posteriormente, todas las unidades de registro deben usar esta variable derivada cuando publiquen tabulados.

Errores de modelo

Cuando se usa un modelo para estimar los valores de variables derivadas (véase sección 6.2.3), los errores de modelo pueden considerarse aleatorios. Los errores de modelo deben examinarse con encuestas especiales, cuyos resultados pueden servir para estimar errores de modelo sistemáticos y aleatorios.

En la sección 6.2.3 presentamos la tabla que aparece a la izquierda en el gráfico 10.2. En la Encuesta sobre Fuerza Laboral (EFL), se entrevistó a 31 458 personas y se les clasificó como empleados o no empleados. Si confiamos en la calidad de la EFL, las probabilidades de los dos tipos de errores de modelo o clasificación de la variable derivada *empleado* en el Registro de Empleo se pueden estimar como 5 y 13%, respectivamente.

Gráfico 10.2 Errores de clasificación en el Registro de Empleo, 1993

Núm. de personas en los datos de prueba	Estimación en el Registro de Empleo		
	Empleados	No empleados	Total
Empleados EFL	22 360	1 158	23 518
No empleados EFL	1 068	6 872	7 940
Total	23 428	8 030	31 458

Errores de clasificación estimados		
Empleados	No empleados	Total
95%	**5%**	100%
13%	87%	100%

Si suponemos que todos los empleados y no empleados están clasificados con estos riesgos de error en el registro, podemos estimar los componentes de calidad de la siguiente manera, donde el número de personas clasificadas como empleados equivale a la suma de dos variables estocásticas con distribuciones binomiales.

En el gráfico 10.3, los verdaderos números de personas empleadas y no empleadas en dos dominios se comparan con los números esperados correspondientes y los errores estándares según los errores de clasificación estimados en el gráfico 10.2.

Gráfico 10.3 Calidad de las estimaciones para dos dominios en el Registro de Empleo

Dominio	Número **verdadero** de Empleados	Número **verdadero** de No empleados	Núm. de personas en el dominio	Número **esperado** de empleados en el registro	Error estándar de la estimación del registro
1	100	100	200	95 + 13 = 108	4.0
2	50	10	60	47.5 + 1.3 = 48.8	1.9

Se espera que una tasa de empleo estimada de 54% (=108/200) en el domino 1 con 200 personas presente un error sistemático de aproximadamente 8/200 = 4 puntos porcentuales, con un error estándar de 4/200 = 2 puntos porcentuales.

Se espera que una tasa de empleo estimada de 81.3% (=48.8/60) en el dominio 2 con 60 personas presente un error sistemático de aproximadamente –1.2/60 = –2 puntos porcentuales, con un error estándar de 1.9/60 = 3 puntos porcentuales.

La información sobre el error sistemático podría usarse para producir estimaciones ajustadas y la información sobre el error estándar para indicar que las celdas de la tabla se basan en observaciones muy escasas.

La conclusión general de esta sección es que debemos conocer la calidad de las variables del registro. Los responsables del registro deben saber qué tan detallados pueden ser los tabulados. En la sección 8.2 mencionamos la sobrecobertura del Registro de Población sueco. La subcobertura global es muy pequeña, pero para algunas categorías de nacidos fuera de Suecia la subcobertura alterará los tabulados detallados. Todos los usuarios del Registro de Población deben conocer esta información sobre la calidad. Algunas variables tienen valores perdidos y los errores de imputación generarán errores aleatorios que alterarán las estimaciones en tabulados demasiado detallados. Asimismo, todos los usuarios de esas variables deben conocer la magnitud de los valores perdidos.

10.2.3 Variación aleatoria en los datos del registro

Cuando interpretamos estimaciones de una encuesta por muestreo, siempre debemos responder las siguientes preguntas:

- ¿Las diferencias o cambios son importantes desde la perspectiva del tema? El usuario debe poder responder esta pregunta sin apoyo.
- ¿Las diferencias o cambios son estadísticamente significativos en comparación con los errores estándares? Los usuarios necesitan ayuda para discernirlo.

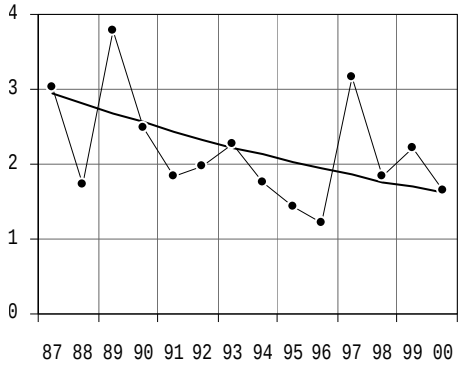
La primera pregunta también se debe hacer cuando se interpretan estimaciones de encuestas basadas en registros, pero no hay una norma similar para la segunda pregunta. Existe el riesgo de que los usuarios “sobreinterpreten” las estimaciones de los censos y las encuestas basadas en registros: como no hay error de muestreo, todo es estadísticamente significativo. Para evitar esta interpretación de las estadísticas basadas en registros, las diferencias y los cambios deben valorarse contra una medida de aleatoriedad en los datos que sea distinta del

error de muestreo. Como la mayoría de las encuestas basadas en registros son anuales o mensuales, recomendamos usar para ese propósito el ruido de las series de tiempo. Con un programa de cómputo para el análisis de *series de tiempo*, una serie mensual se puede descomponer en tendencia, estacionalidad y ruido. Del mismo modo, en el caso de una *serie de tiempo* anual se puede descomponer en tendencia y ruido. Incluso cuando no hay error de muestreo, puede haber un nivel sustancial de ruido en la serie de tiempo. En el análisis de series de tiempo esto se suele interpretar como una *variación aleatoria natural*.

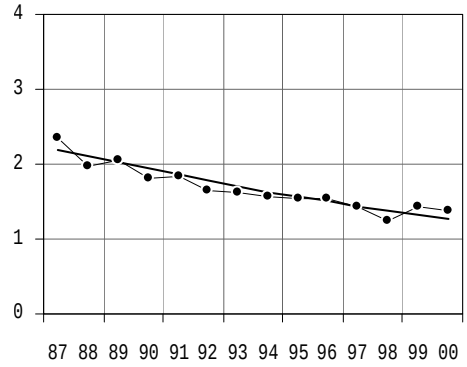
En los gráficos siguientes se presentan frecuencias reales de accidentes. Las variaciones de corto plazo en estas frecuencias no se deben interpretar como indicios de que han cambiado los riesgos subyacentes de accidentes. Mientras que en las regiones pequeñas con poca población infantil la variación aleatoria natural será extremadamente marcada, en las regiones grandes con mucha población infantil será menor.

Gráfico 10.4 Accidentes en carretera, niños de 0 a 12 años, valores reales y tendencias estimadas. Número de lesionados por cada 1 000 niños/año

A. Condado con una población aproximada de 5 000 niños



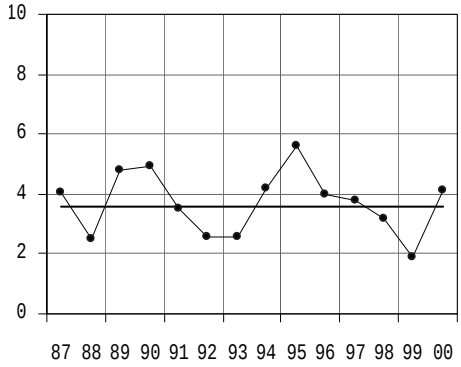
B. Condado con una población aproximada de 125 000 niños



En los gráficos C y D siguientes se muestran otras dos pautas de tendencias. Las tendencias de estos gráficos se seleccionaron de modo que los residuos de las series de tiempo (el "ruido") no se autocorrelacionen y su error estándar se acerque a la desviación estándar teórica de acuerdo con la distribución de Poisson.

Gráfico 10.4 (Continuación) Número de lesionados por cada 1 000 niños/año

C. Condado sin tendencia



D. Condado con tendencia con promedio móvil

Las series de los gráficos A a D se crearon a partir de estadísticas basadas en el Registro de Pacientes y el Registro de Población suecos. El mensaje para los usuarios es claro:

En el gráfico C no hay cambios estadísticamente significativos, mientras que en el gráfico D el riesgo de accidentes se incrementó en el periodo 1987-1994, aunque posteriormente ha ido a la baja. Recomendamos analizar las series de tiempo mensuales, trimestrales o anuales de las encuestas basadas en registros y usar el error estándar de los residuos de las series de tiempo para evaluar si las diferencias y los cambios son estadísticamente significativos.

10.2.4 Errores de medición

Hay muchas diferencias importantes entre las estructuras de error encontradas en los registros estadísticos basados en datos administrativos y las encontradas en los datos captados en encuestas por muestreo. Ya mencionamos que los errores de integración son importantes para las encuestas basadas en registros y que este tipo de errores no existe en las encuestas por muestreo, donde todas las variables se captan mediante cuestionarios o entrevistas. A su vez, en las encuestas por muestreo predominan los errores de muestreo, que no existen en las encuestas basadas en registros.

Los procesos de medición también difieren bastante entre estos tipos de encuesta. En las encuestas por muestreo basadas en cuestionarios o entrevistas, los procesos cognitivos para responder las preguntas se modelan como (véase Groves *et al*. (2004), p. 202):

– Comprensión de la pregunta
– Recuperación de la información
– Evaluación y estimación
– Informar una respuesta

¿Serán relevantes los mismos procesos cognitivos cuando las personas suministran datos a una autoridad administrativa, ya sea como particulares o como representantes de una empresa? Estos procesos cognitivos o psicológicos también ocurren en la notificación de datos administrativos, pero no pensamos que sean tan importantes. En cambio, las reglas administrativas y la legislación son factores importantes y cuando se presentan datos de empresas los principios y prácticas contables tienen mayor peso que la psicología.

En el siguiente gráfico comparamos ambas formas de recopilar datos:

Gráfico 10.5 Errores de medición: comparación de métodos para la captación de datos

Captación de datos en encuestas por muestreo	Captación de datos en sistemas administrativos
Estructura subyacente de la pregunta: *¿Podría tratar de entender nuestras preguntas y hacer memoria?* *No es necesario que responda y no importa lo que responda, no le haremos ningún daño.*	Estructura subyacente de la pregunta: *1. Declare la facturación del mes anterior antes del día 12 de este mes.* *2. Pague 25% de la facturación declarada antes del día 12 de este mes.* *3. En caso de no declarar y pagar, tendrá que pagar más.*
Cuestionario para personas: ¿Se entrevista a la persona indicada dentro del hogar?	Suministro de datos de personas a las autoridades: en nuestra declaración anual de impuestos, sólo agregamos nuestra firma y tal vez intentemos hacer algunas deducciones.
Cuestionario para empresas: ¿Se entrevista a la persona indicada dentro de la empresa?	Suministro de datos de empresas a las autoridades: es un deber habitual de los profesionales, el sistema contable de la empresa emite el informe correspondiente. Los errores son errores ya sea en las rutinas contables, la captura o la digitalización.

El efecto de los entrevistadores puede resultar perturbador.	No hay entrevistadores, por lo tanto no hay efectos de los entrevistadores.
En las investigaciones de mercado las preguntas que inducen la respuesta suelen ser un problema.	Cuestiones jurídicamente complejas.
Variables recopiladas mediante un cuestionario o una entrevista en las encuestas por muestreo (o censos): los errores de medición son importantes.	Las variables estadísticas de las encuestas basadas en registros muchas veces son variables derivadas de variables administrativas: los errores de relevancia y los errores de modelo son importantes.

Cabe señalar que algunas de las preguntas o variables recopiladas en un sistema administrativo revisten importancia jurídica, mientras que otras preguntas no. La calidad de estas preguntas no importantes puede ser menor: podemos responder lo que queramos, no nos afectará en nada; se aplican las mismas condiciones que a una pregunta de una encuesta por muestreo.

10.3 LA VISIÓN DE LOS USUARIOS Y LOS PRODUCTORES DE ESTADÍSTICAS RESPECTO A LA CALIDAD

¿Cómo se usan las estadísticas basadas en registros? ¿Cuáles son las demandas de los usuarios en lo relativo a la calidad? Biemer y Lyberg (2003) abordan el tema de la calidad en las encuestas por muestreo. El punto de partida es *una* estimación, cuyo error total se divide en doce componentes. Los errores que ocurren en diferentes etapas de una encuesta por muestreo pueden ser aleatorios o sistemáticos. Los *errores aleatorios* vuelven inciertas las estimaciones, pero no causan distorsiones. En cambio, los *errores sistemáticos* sí provocan distorsiones, es decir, el valor que se busca se sobrestima o subestima.

Gráfico 10.6 Riesgo de errores aleatorios y sistemáticos por fuente de error principal

Fuente de error	Riesgo de errores aleatorios	Riesgo de errores sistemáticos
Error de especificación	Bajo	Alto
Error de marco	Bajo	Alto
Error de no respuesta	Bajo	Alto
Error de medición	Alto	Alto
Error de procesamiento de datos	Alto	Alto
Error de muestreo	Alto	Bajo

Fuente: Biemer y Lyberg (2003, p. 59).

Ésta es la visión sobre calidad que tiene el *productor* de estadísticas, para quien es fundamental saber qué partes de la encuesta funcionan bien y cuáles no. A partir de ese conocimiento, se pueden mejorar los procesos donde se presentan los errores más graves. Ahora nos referiremos a los componentes de los errores correspondientes en las encuestas basadas en registros.

Platek y Särndal (2001) analizan la calidad de las estadísticas oficiales desde el punto de vista del *usuario*, quien está interesado en las respuestas a preguntas como: "¿puedo confiar en estas estadísticas?" y "¿son adecuadas para mis propósitos?". Los usuarios quieren una garantía de calidad. ¿Pueden dársela los estadísticos? ¿Qué forma tendría?

En muchas ocasiones, las respuestas que las oficinas de estadística dan a estas preguntas no son lo suficientemente claras. Platek y Särndal afirman que la calidad de los datos tiene un

significado distinto para las diferentes categorías de funcionarios en una oficina de estadística:

– para los metodólogos estadísticos es una cuestión de exactitud;
– para los especialistas en los distintos temas tiene que ver con el contenido y la presentación;
– para los especialistas en informática se refiere al funcionamiento eficiente de los sistemas y el procesamiento de datos;
– para los administrativos es una cuestión de funcionamiento de los presupuestos y los plazos.

Del conocimiento detallado de la calidad a un panorama general de la calidad

En esta sección nos centraremos en la diferencia entre un conocimiento profundo y un panorama general. Las distintas categorías de funcionarios cuentan con un conocimiento sólido y profundo de una gran diversidad de factores que afectan la calidad. El metodólogo piensa en función de las diferentes fases de la encuesta y todas las fuentes de error que pueden conllevar, mientras que el especialista en informática piensa en el sistema de producción y todos los posibles errores de procesamiento. Ambos tienen un vasto conocimiento, pero tal vez no el panorama general que requiere el usuario.

Los especialistas en los diversos temas suelen tener el contacto más estrecho con los usuarios de las estadísticas, por lo que son los más indicados para ofrecer este panorama general de la calidad. Sin embargo, para ello deben colaborar estrechamente con los metodólogos y los especialistas en informática, además de ser sensibles a las necesidades de los usuarios.

El conocimiento detallado de los diferentes componentes de calidad se adquiere siguiendo los lineamientos para el aseguramiento de la calidad en relación con los procesos de producción de los registros estadísticos. Partiendo de este conocimiento, se puede hacer una valoración general de la calidad del registro y la encuesta basada en registros. Considerando nuestras conclusiones anteriores, adoptaremos dos enfoques distintos para analizar la calidad de las estadísticas basadas en registros:

– En primer lugar, describiremos los indicadores de calidad en las diferentes etapas de una encuesta basada en registros. El productor de las estadísticas basadas en registros debe usar estos indicadores para fines de aseguramiento de la calidad en las distintas etapas del proceso de producción.
– Por último, describimos las valoraciones generales de la calidad que son comprensibles para los usuarios de las estadísticas basadas en registros.

10.4 CONOCIMIENTO DETALLADO DE LAS CARACTERÍSTICAS DE UN REGISTRO

El propósito del aseguramiento de la calidad consiste en investigar y corregir los defectos en la calidad, lo que puede ocurrir en diferentes partes del trabajo con un registro. El modelo de error tradicional acompaña las distintas etapas de una encuesta por muestreo. También describimos la calidad de las estadísticas basadas en registros apoyándonos en conjuntos de indicadores para las diversas fases de generación de un registro:

Gráfico 10.7 Las fases para crear un registro: características de calidad correspondientes

Fases para crear un registro	Características de la calidad de un registro
1. Determinación de los objetivos de investigación 2. Inventario 3. Planeación	Grupo 1: Competencia en el tema y capacidad de desarrollo
4. Contacto con los proveedores, recepción de los datos administrativos	Grupo 2: Contacto con los proveedores y validación
5a. Integración: el conjunto de objetos	Grupo 3: Características del conjunto de objetos, tanto en el registro base como en términos generales
5b. Integración: el tipo de objetos	Grupo 4: Características de la definición de los objetos
5c. Integración: las variables	Grupo 5: Características de las variables
6. Documentación	Grupo 6: Características de la documentación

10.4.1 Fase de determinación de los objetivos de investigación: su efecto en el registro

¿Qué necesidades y requisitos estadísticos se supone que debe satisfacer el registro? ¿Se ha establecido contacto con usuarios especializados o investigadores? ¿La oficina de estadística elaborará sus propios análisis e informes especializados? Si se sabe lo que se requiere de un registro, será posible hacerse una idea clara de los usos que podrían dársele. Los usuarios especializados de los productos estadísticos constituyen un grupo importante cuando se considera la creación y aplicación de un registro, ya que a menudo tienen experiencias e ideas valiosas que conviene documentar.

10.4.2 Fase de inventario: ¿cómo ha afectado al registro?

En la fase de inventario se evalúan las distintas fuentes utilizadas para producir el nuevo registro. Pueden ser fuentes administrativas y registros estadísticos que ya existen en el sistema. En su conjunto, ¿tienen esas fuentes un contenido valioso? Si se han integrado varias fuentes al registro, esto representará una ventaja para los nuevos usuarios. En esta fase también se debe investigar si hay otras fuentes relacionadas con el área de estudio que no se han utilizado. Ésta puede ser una oportunidad para una búsqueda activa de fuentes nuevas.

10.4.3 ¿Se planean cambios?

¿Hay planes de cambios o mejoras en relación con la población de registro, las definiciones de los objetos o el contenido de las variables? De ser así, es una señal tanto de defectos en el registro actual como de un trabajo de desarrollo encaminado a aumentar la usabilidad del registro. Podemos sintetizar las secciones 10.4.1 a 10.4.3 de la siguiente manera:

Grupo 1. Competencia en el tema de estudio y capacidad de desarrollo

Indicador de calidad	Implica	¿Cómo se mide o valora?	Véase sección
1a Buen contacto con usuarios	Competencia en el tema de estudio	Valoración cualitativa	10.4.1
1b Realización de análisis propios			
1c Integración de fuentes	Intención de uso diversificado		10.4.2
1d Búsqueda de nuevas fuentes	Capacidad de desarrollo		10.4.3
1e Planeación de cambios			

10.4.4 Contactos con proveedores y validación

¿Cómo es el contacto con los proveedores de datos? Este contacto debe ser lo suficientemente profundo para que el proveedor entienda las necesidades del producto estadístico y el receptor, las condiciones que rigen el sistema administrativo. El contacto se debe renovar con regularidad, para que el receptor obtenga información sobre futuros cambios.

¿Cómo se reciben los datos y metadatos del proveedor? Se debe examinar la validación de estos datos: ¿sólo habrá una validación simple de cada fuente administrativa o también una validación de la consistencia cuando se comparan varias fuentes? También se pueden analizar las experiencias adquiridas durante el proceso de validación de datos. ¿Se ha documentado la frecuencia de los diversos tipos de error? ¿Cómo se retroalimenta al proveedor con esta información? Sintetizamos estas consideraciones en el siguiente cuadro:

Grupo 2. Indicadores sobre el funcionamiento del contacto con proveedores y la validación

Indicador de calidad	Implica	¿Cómo se mide o valora?	Véase sección
2a Grupo de la oficina de estadística regularmente en contacto con el proveedor o una persona ocasionalmente en contacto	Conocimiento del tema de estudio amplio o limitado, respectivamente	Valoración cualitativa	6.3.1 6.3.5
2b Identificación de registros no válidos y duplicados durante la validación	Fallas de calidad en la fuente administrativa	Número o proporción de elementos no válidos o duplicados	6.3.4
2c Identificación de diferencias entre la población de registro y el registro base	Subcobertura Sobrecobertura	Número o proporción de elementos perdidos y extraños	
2d Identificación de coincidencias falsas durante la validación	Vínculos que indican una relación mala o una identidad incorrecta	Número o proporción de coincidencias falsas	
2e Identificación de objetos o valores de variables perdidos durante la validación	Entrega incompleta	Número o proporción de objetos o valores de variables perdidos	
2f Identificación de valores de variables obviamente incorrectos durante la validación	Defectos de calidad en la fuente administrativa	Número o proporción y magnitud de los diferentes tipos de errores	
2g Validación de cada fuente y validación de la consistencia entre las fuentes, comparando varias fuentes	Intención de uso diversificado	Valoración cualitativa	6.3

10.4.5 Fase de integración: ¿cómo ha sido creado el conjunto de objetos?

En esta fase, se debe analizar la manera en que se han procesado las fuentes con que se cuenta para asegurar que el nuevo registro contenga el conjunto de objetos deseado. Esto conlleva la validación de los datos administrativos, el emparejamiento de distintas fuentes y la selección de objetos o el procesamiento de referencias temporales a fin de producir el conjunto de objetos para el momento o el periodo designado.

En los siguientes dos cuadros presentamos indicadores para registros base y los que se pueden aplicar en general a los registros estadísticos.

Grupo 3.1 Indicadores para las características del conjunto de objetos en un registro base

Indicador de calidad	Implica	¿Cómo se mide o valora?	Véase sección
3.1a Retraso en la presentación de datos	La sobrecobertura y la subcobertura en la versión actual del registro generan errores en los datos transversales, así como cambios en los errores a lo largo del tiempo	Se compara la versión actual del registro con una versión basada en todas las fuentes disponibles cuando se cuenta con la información completa. Se indica la magnitud de las correcciones, la duración del retraso y la frecuencia de los cambios.	Más adelante
3.1b Defectos en la presentación de datos respecto a los objetos que han dejado de existir	Sobrecobertura, que genera errores en los datos transversales, así como cambios en los errores a lo largo del tiempo	Esto puede causar valores inverosímiles en las tablas. Se usan las encuestas existentes con captación de datos propia: correo devuelto, sin contacto para entrevistas telefónicas. Se hacen comparaciones con la información de otros registros y se estima la magnitud de la sobrecobertura.	8.2
3.1c Objetos con información incorrecta	Sobrecobertura y subcobertura	Se investiga el número o la proporción de datos de notificación incorrectos que se corrigieron o anularon.	Más adelante

Retraso en la presentación de datos

En lo concerniente al Registro de Población sueco, la intención es dar seguimiento continuo a distintos indicadores de calidad mensuales sobre retrasos en el registro de nacimientos, defunciones y cambios de domicilio. Las diferentes autoridades hacendarias locales pueden presentar diferentes patrones de retraso.

Por su parte, el Registro Empresarial tiene la finalidad, por un lado, de dar seguimiento mensual a las frecuencias de cambio respecto a las empresas recién registradas y dadas de baja, junto con las frecuencias relacionadas con el domicilio y la rama industrial, y, por otro lado, monitorear qué tan actualizada está la información.

Información incorrecta sobre objetos

En el Registro de Población sueco, hay una base de datos de calidad especial donde se guarda toda la información sobre notificaciones incorrecta que el Consejo Nacional Tributario ha corregido o anulado posteriormente.

Grupo 3.2 Indicadores de las características del conjunto de objetos en general

Indicador de calidad	Implica	¿Cómo se mide o valora?	Véase sección
3.2a El registro de la población no concuerda con el registro base	Falta de consistencia y coherencia	Se empareja el registro con el registro base y se analizan los no emparejamientos.	5.4.4, 5.4.5, capítulo13
3.2b La fuente administrativa contiene objetos no presentes en el registro base	Subcobertura en el registro base	Se analiza la magnitud y la causa del no emparejamiento, y se notifica al registro base.	5.4.7
3.2c El registro base contiene objetos no presentes en la fuente	Subcobertura en el registro fuente o sobrecobertura en el registro base	Se analiza la magnitud y la causa del no emparejamiento.	5.5.7

3.2d	Errores de emparejamiento: falsas coincidencias	Relación incorrecta, no hay correspondencia entre los objetos aunque tengan los mismos valores en las variables de vinculación	Se comparan las variables del registro integrado.	5.4.7 5.5.6 6.3.2

El registro de la población no concuerda con el registro base
Cuando estudiamos el sistema de registros de la Oficina de Estadística de Suecia identificamos un ejemplo de poblaciones inconsistentes en las estadísticas agropecuarias, donde las empresas de la rama no eran idénticas a las encontradas en el Registro Empresarial. Ocurría lo mismo con las empresas dedicadas a la educación y la energía. Otro ejemplo es el de las estadísticas sobre establecimientos de hospedaje, donde la población de empresas se conjuntó independientemente del Registro Empresarial.

10.4.6 Fase de integración: ¿cómo han sido definidos los objetos?

¿Qué procesamiento se ha llevado a cabo para verificar y cambiar las definiciones de objetos? Por ejemplo, ¿los datos administrativos se han verificado y ajustado para que las definiciones sean las requeridas? ¿Se han formado objetos derivados en el nuevo registro? ¿Se verifica la calidad de las definiciones de los objetos? ¿Se llevan a cabo cuestionarios de mantenimiento o evaluaciones del registro? ¿De qué magnitud son los errores?

Grupo 4. Indicadores para las características de las definiciones de los objetos

Indicador de calidad		Implica	¿Cómo se mide o valora?	Véase sección
4a	La definición de los objetos se desvía de la norma	Falta de consistencia y coherencia	Valoración cualitativa	Más adelante
4b	La definición de los objetos no es comparable a lo largo del tiempo	Falta de consistencia y coherencia	Valoración cualitativa	6.4
4c	¿Se llevan a cabo cuestionarios de mantenimiento del registro?	Conciencia sobre la calidad	¿Con qué frecuencia se aplican esos cuestionarios y en relación con qué categorías de objetos?	5.4.6–7 5.5.3 6.3.2

La definición de los objetos se desvía de la norma
En el Registro de Evaluación Inmobiliaria de Suecia, los objetos (y también las variables) se definen en términos administrativos y no en coordinación con otras encuestas de la Oficina de Estadística de Suecia; el término “bloque de departamentos”, por ejemplo, se define de manera distinta. De igual modo, en el Registro de Población la definición de una familia desde el punto de vista tributario no concuerda con las definiciones de hogar de otras encuestas de la Oficina de Estadística de Suecia.

10.4.7 Fase de integración: ¿cómo han sido creadas las variables?

En esta fase se debe evaluar cuál fue el proceso para producir las variables deseadas. Esto puede incluir considerar si las variables de las fuentes administrativas se han validado y también el alcance y tratamiento de los valores perdidos. También se deben documentar las diversas fuentes de donde se han importado las variables.

Asimismo, se debe investigar el alcance de todos los errores posibles, como errores de medición o de clasificación en las variables de expansión. Los métodos aplicados para detectar errores podrían incluir encuestas por muestreo o encuestas de evaluación

especiales. También podrían realizarse grupos de enfoque y entrevistas cognitivas para analizar las formas administrativas e identificar fuentes de errores de medición.

Grupo 5. Indicadores de las características de las variables

Indicador de calidad	Implica	¿Cómo se mide o valora?	Véase sección
5a Errores de emparejamiento: coincidencias falsas	Error en la variable de vinculación	Hay objetos idénticos con diferentes valores para las variables de vinculación.	Más adelante
5b Errores de emparejamiento: falsas coincidencias a pesar de ser los mismos valores para las variables de vinculación	Error en la variable de vinculación	Relación incorrecta, no hay correspondencia entre los objetos aunque tengan los mismos valores en las variables de vinculación. Se deben comparar las variables del registro integrado: las relaciones inverosímiles indican errores.	5.4.7 5.5.5 5.5.6 6.3.2
5c La definición de las variables se desvía de la norma	Falta de consistencia y coherencia	Valoración cualitativa	Más adelante
5d Valores perdidos y errores de imputación	Defectos de calidad de la fuente administrativa	Se identifican la magnitud y las causas de los valores perdidos para cada variable.	6.1.2 6.3 8.1
5e Retraso en la presentación de los datos	Error de revisión	Se comparan las estimaciones preliminares con las definitivas.	Más adelante
5f Valores preliminares incorrectos	Error de revisión	Se investiga el número o la proporción de entradas incorrectas que se han corregido o anulado.	Más adelante
5g Variable no usada por el proveedor	Error de medición	Se investiga si el proveedor usa y valida la variable.	Más adelante
5h Estimaciones sospechosamente altas o bajas	Error de medición	Estos errores se pueden evaluar por medio de encuestas por muestreo y entrevistas cognitivas.	Más adelante
5i Errores de codificación	Error de clasificación	Los mismos datos se codifican de nuevo de manera independiente para estimar los errores de codificación.	6.2.4
5j Errores de agregación en variables de valores múltiples	Error en datos transversales y a lo largo del tiempo	El registro sólo contiene el valor más importante de la variable.	9.2
5k Error de clasificación en las variables de expansión	Error estructural, relación debilitada entre variables	Encuestas de control de calidad. Los errores se pueden evaluar y estimar por medio de encuestas por muestreo especiales.	Más adelante
5m Error de modelo en la variable derivada	Error de medición o de clasificación	Los errores se pueden evaluar y estimar por medio de encuestas por muestreo tradicional o especial.	6.2.3

Error de emparejamiento: coincidencias falsas

Los domicilios de los establecimientos son un vínculo entre el Registro Empresarial y el Registro Inmobiliario suecos. Un intento de emparejar los establecimientos por medio de este vínculo falló en alrededor de 30% de los casos de las personas empleadas en el municipio, donde el emparejamiento fue menos exitoso. Al validar todos los establecimientos con al menos 25 empleados, este porcentaje podría reducirse a alrededor de 10 por ciento.

La definición de las variables se desvía de la norma
Las definiciones en el Registro de Comercio Exterior sueco no son las mismas que para las Cuentas Nacionales y la Encuesta Empresarial Estructural.

Retraso en la presentación de los datos (i)
Los retrasos en la presentación de los datos mensuales de la industria de la construcción habitacional sueca ocasionaron una subestimación del número de departamentos en construcción. El error se corrigió en el momento de la revisión de las cifras preliminares. Es importante dar seguimiento a la duración del retraso y hacer estimaciones actualizadas del error corregido en la revisión, incluyendo sus componentes sistemáticos y aleatorios.

Retraso en la presentación de datos (ii)
Los retrasos en la presentación mensual de datos sobre sumas salariales, impuesto laboral, etc. dan origen a una no respuesta. Las estimaciones preliminares se basan en valores imputados y se corregirán cuando se reciba la información.

Valores preliminares incorrectos
La información de los contribuyentes podría modificarse, ya sea porque los propios contribuyentes indican tales modificaciones o porque las autoridades hacendarias lo hicieron.

Información no utilizada por el proveedor de datos
En Suecia, el empleador debe indicar en el estado de resultados el primer y el último mes de empleo. Las autoridades hacendarias no utilizan esta información, por lo que es de mala calidad.

Estimaciones sospechosamente altas o bajas
Los propietarios de casas solas en Suecia tienden a subestimar el espacio residencial cuando se trata de una valuación fiscal porque así pueden pagar menos impuesto predial.

Errores de clasificación en la variable de expansión
La estructura de la población, o la distribución de acuerdo con la variable de expansión (por ejemplo, la rama industrial), es incorrecta. Las correlaciones están debilitadas y las comparaciones entre categorías, distorsionadas. Supongamos, por ejemplo, que se hará una comparación entre empresas manufactureras y de servicios desde el punto de vista de la rentabilidad. Si no hay error de clasificación en la rama industrial, la comparación es correcta, como en el gráfico 10.8, mientras con un error de clasificación de 10% la comparación aparecerá como en el gráfico 10.9.

Gráfico 10.8 Sin error de clasificación

Rama industrial	Facturación	Utilidades	Rentabilidad
Manufactura	300	30	10%
Servicios	250	10	4%

Gráfico 10.9 Con error de clasificación

Facturación	Utilidades	Rentabilidad
300–30+25=295	30–3+1=28	9,5%
250–25+30=255	10–1+3=12	4,7%

10.4.8 Documentación como parte del aseguramiento de la calidad

El trabajo de documentación también es una parte importante del aseguramiento de la calidad. El uso incorrecto o indiscriminado de los datos administrativos puede evitarse utilizando metadatos que proporcionen información sobre los problemas de comparabilidad. Los cambios en el sistema administrativo ocasionan este tipo de problemas y, por lo tanto, deben documentarse. De lo contrario, existe el riesgo de sacar conclusiones erróneas.

Como es posible que varios usuarios que trabajan con el sistema de registros utilicen un registro estadístico, los registros se deben documentar de tal forma que sean accesibles para todos.

Los metadatos desempeñan un papel fundamental en el trabajo sobre estadísticas basadas en registros. En el caso de la integración de diferentes registros, es necesario conocer las definiciones y cuáles son los problemas de comparabilidad. También es importante documentar los métodos de procesamiento a fin de facilitar la formulación de una metodología y el intercambio de experiencias.

En el capítulo 11 veremos la documentación de registros y encuestas basadas en registros. Una buena documentación debe dar respuesta a las preguntas del siguiente cuadro.

Grupo 6. Indicadores para las características de la documentación

Indicador de calidad	Implica	¿Qué se debe documentar?	Véase sección
6a Definición del registro de la población	Contenido de las estadísticas	¿La definición está completa? Tipo de objeto, objetos incluidos, referencias temporales, definición geográfica, relación de objetos con área. ¿Qué fuentes se han integrado? ¿Qué métodos de procesamiento se han usado para crear la población de registro?	5.4.1
6b Definición de las variables	Contenido de las estadísticas	¿La definición está completa? Tipo de objeto, referencia temporal de la variable, método de medición, escala. ¿Qué fuentes se han integrado? ¿Qué métodos de procesamiento se han usado para crear las variables del registro?	6.1.1
6c Población comparable a lo largo del tiempo	Calidad de las series de tiempo	Cambios de nivel en las series de tiempo y sus efectos. ¿Qué procesamiento se ha hecho para permitir la comparabilidad a lo largo del tiempo?	8.3 9.3
6d Variables comparables a lo largo del tiempo	Calidad de las series de tiempo	Cambios de nivel en las series de tiempo y sus efectos. ¿Qué procesamiento se ha llevado a cabo para permitir la comparabilidad a lo largo del tiempo?	8.3 9.3
6e Objetos de la población comparables a lo largo del tiempo	Calidad longitudinal	¿Los objetos tienen identidades estables? ¿Se les puede dar seguimiento a lo largo del tiempo? ¿Qué procesamiento se ha llevado a cabo para permitir la comparabilidad a lo largo del tiempo a escala de los objetos?	6.4
6f Valores de variables de los objetos comparables a lo largo del tiempo	Calidad longitudinal	¿La variable puede describir los cambios a escala de los objetos?	6.4

10.4.9 Indicadores de calidad: conclusiones

A menudo los responsables de un registro han adquirido conocimientos cualitativos sobre la calidad del registro, gracias a su conocimiento sobre el tema de que se trate, sus observaciones y su experiencia. Esto significa que *están familiarizados con la ocurrencia de varios tipos de errores, pero no pueden medir su frecuencia o grado de importancia*.

Este tipo de conocimientos también se debe documentar, tanto para advertir a los usuarios sobre un uso indiscriminado como para indicar dónde se requieren mejoras. Lindström (1999) propone la siguiente escala de conocimiento sobre diferentes aspectos del registro:

Gráfico 10.10 Niveles de conocimiento sobre calidad

1. Ningún conocimiento
2. Información cualitativa vaga basada en la apreciación
3. Documentación de los procesos de producción y sus características
4. Información cualitativa sistemática basada en encuestas
5. Indicadores cuantitativos, por ejemplo, magnitud del no emparejamiento
6. Mediciones cuantitativas de la calidad de la estimación

Los grupos de indicadores presentados en los cuadros de las páginas anteriores pueden usarse para describir las características del registro en cualquiera de los niveles 2 a 6.

¿Qué nos dicen estos indicadores sobre la calidad del registro?

Los 40 indicadores de los grupos 1 a 6 ofrecen una referencia útil a la hora de decidir cómo investigar y describir las características de determinado registro. Dado que las condiciones (es decir, las fuentes administrativas y las posibilidades de usar el sistema de registros) difieren de un registro a otro, también la importancia de los indicadores diferirá de un registro a otro.

No se trata de usar todos los indicadores para todos los registros, sino de elegir aquellos que son relevantes para el registro de que se trate y usarlos para describir la calidad del registro.

Muchos de los indicadores de calidad mencionados sólo se aplican a registros y encuestas basadas en registros, no a encuestas con una captación de datos propia.

10.5 VALORACIÓN GENERAL DE LA CALIDAD

En ocasiones, las estadísticas basadas en registros se pueden usar mediante el cálculo de un valor resumen:

El empleo en la industria manufacturera de nuestro municipio se ha incrementado 1.6 por ciento.

Sin embargo, muchas veces se estudian grandes tablas con cientos de estimaciones a fin de descubrir los patrones de las relaciones y de las trayectorias de las series de tiempo. Los patrones identificados dan pie a conclusiones:

La estructura laboral de nuestro municipio ha empeorado en cuanto a edad y nivel de escolaridad, lo que perjudica el sector de alta tecnología.

Esta conclusión se basa en muchas comparaciones con otros municipios a lo largo del tiempo. En este caso resulta imposible hacer un análisis basado en los errores numéricos de los cientos de estimaciones en los que se basa la conclusión. Como cada error se puede dividir en muchos componentes que se deben evaluar de manera independiente, la situación se vuelve sumamente complicada. Lo que el usuario quiere saber es si es posible llegar o no a la conclusión de que se trate. ¿Hay complicaciones estadísticas que el usuario no ha considerado?

Sobre la base del conocimiento detallado de las características de un registro, se puede hacer una valoración general de la calidad tanto del registro como de la encuesta basada en registros.

Calidad de los registros y las encuestas basadas en registros

La calidad del *registro* se debe describir en términos generales, de modo que los usuarios en potencia puedan determinar si conviene a sus propósitos. La descripción debe

relacionarse con las diversas áreas de aplicación que pueden resultar de interés. Distinguimos tres maneras de usar registros y sus aspectos cualitativos correspondientes:

– *Calidad de corte transversal:* ¿qué comparaciones se pueden hacer dentro del registro?
– *Calidad de las series de tiempo:* ¿qué comparaciones se pueden hacer a lo largo del tiempo en el nivel agregado?
– *Calidad longitudinal:* ¿qué comparaciones se pueden hacer a lo largo del tiempo en el nivel micro?

La calidad de una *encuesta basada en registros* se debe describir para *un* uso particular del registro. ¿La calidad de las estimaciones es buena o mala para el uso deseado? Se debe describir la relevancia y exactitud de las estimaciones. A continuación se comparan distintas maneras de describir la calidad general:

Gráfico 10.11 La calidad general de los registros y las encuestas basadas en registros

Calidad	Registro	Encuesta basada en registros específica
Relevancia	Sólo se dan las definiciones	¿Las definiciones son adecuadas y funcionales? Esto se analiza en detalle.
Calidad de corte transversal	¿Qué comparaciones se pueden hacer?	La calidad sólo se describe para el uso específico de que se trate. ¿La calidad es buena o mala? Se describe la calidad de las estimaciones.
Calidad de las series de tiempo	¿Qué comparaciones se pueden hacer?	
Calidad longitudinal	¿Qué comparaciones se pueden hacer?	

10.5.1 Una valoración general de la calidad de un registro

En una valoración general de la calidad de un registro estadístico se deben abordar los siguientes aspectos:

– Las definiciones de la población de registro y las variables utilizadas deben estar disponibles y ser fáciles de entender.
– ¿Qué comparaciones se pueden hacer dentro del registro?
  Ejemplo: Si, en el área de estadísticas regionales, se quieren comparar municipios, ¿la no respuesta parcial es, a grandes rasgos, la misma en todos los municipios o se pueden hacer comparaciones mediante valores imputados?
– ¿Qué comparaciones se pueden hacer a lo largo del tiempo en un nivel agregado?
  Ejemplo: ¿La no respuesta parcial es, a grandes rasgos, la misma para todos los años o se pueden hacer comparaciones a lo largo del tiempo mediante valores imputados? ¿Ha habido cambios en los datos administrativos que dificultan las comparaciones a lo largo del tiempo? ¿El procesamiento del registro se ha llevado a cabo con la intención de facilitar la comparabilidad entre años?
– ¿Qué comparaciones se pueden hacer en el nivel micro a lo largo del tiempo? La posibilidad de hacer comparaciones longitudinales correctas impone los requisitos más estrictos para un registro. ¿Qué verificación o procesamiento se ha llevado a cabo para lograr esto?

Para esta valoración general se requiere un conocimiento profundo del registro, lo que significa que es necesario contar con documentación detallada.

10.5.2 Una valoración general de la calidad de una encuesta basada en registros

Una valoración general de la calidad de una encuesta basada en registros debe dar respuesta a las siguientes preguntas:

- ¿Cómo está definida la *población objetivo* de la encuesta? ¿La definición es adecuada considerando el propósito de la encuesta? ¿Hay diferencias importantes entre la población objetivo y la población de registro?
- ¿Cómo están definidas las *variables* de la encuesta? ¿Las definiciones son adecuadas considerando el propósito de la encuesta? Esto se refiere a las variables de mayor importancia para la encuesta y las principales variables de agregación y expansión en las que se basa el análisis.
- ¿Qué *comparaciones* se pueden hacer en la encuesta? ¿La calidad del registro es suficiente para estas comparaciones?
- ¿Las *estimaciones* se calculan de la mejor manera? Para efectos de la calidad transversal y de las series de tiempo, es importante que las variables de valores múltiples y los valores perdidos se manejen usando métodos de estimación apropiados. En lo que respecta a la calidad de las series de tiempo, también es importante que los cambios de nivel en las series de tiempo estén vinculados.
- ¿Los *resultados* se interpretan de manera razonable a la luz de la incertidumbre de las estimaciones? No se debe considerar que las tablas basadas en los registros estadísticos ofrecen información exacta, pues puede haber variación aleatoria y otras fuentes de error, de modo que es importante tener cuidado de no "sobreinterpretar" los resultados.

Para esta valoración general, debemos saber cuál es la calidad del registro y tener un conocimiento profundo del tema de estudio. Esto significa que la persona que lleva a cabo la valoración debe estar muy familiarizada con los objetivos de investigación de que se trate.

10.6 PRINCIPALES PROBLEMAS DE CALIDAD EN LOS DIFERENTES TIPOS DE ENCUESTAS

Como mencionamos en la sección 10.2, la eficiencia del proceso de captación de datos determina en primera instancia la calidad de la encuesta por muestreo o censo. Esto significa que los errores de muestreo, de medición y de no respuesta son problemas de calidad relevantes. En las secciones 5.4.8 y 5.4.9 abordamos los errores de cobertura. Como las poblaciones marco suelen basarse en fuentes disponibles de manera preliminar y, por lo mismo, menos confiables, los errores de cobertura son un problema más serio en las encuestas basadas en una captación de datos que en las basadas en registros.

En las encuestas basadas en registros, se usan datos administrativos y registros para propósitos estadísticos. Los registros administrativos se procesan de modo que los conjuntos de objetos, los tipos de objetos y las variables satisfagan las necesidades estadísticas. En un registro estadístico, las definiciones de población de registro, objetos y variables determinan los errores de relevancia de las encuestas basadas en registros que usan ese registro. Estos errores de relevancia se pueden evaluar a partir de la documentación o el registro, como se describe en los indicadores de calidad del grupo 6 (sección 10.4.8).

Cuando se crean registros estadísticos, se integran muchas fuentes distintas. La calidad de las variables de vinculación, las fuentes y los métodos utilizados determina los errores de integración del nuevo registro, descritos por los indicadores de calidad de los grupos 3.1, 3.2, 4 y 5 (secciones 10.4.5 a 10.4.7).

A continuación resumimos las consideraciones anteriores. En nuestra opinión, los principales problemas de calidad para las estadísticas basadas en registros son los errores de relevancia y de integración.

Gráfico 10.12 Principales problemas de calidad en los diferentes tipos de encuesta

Encuesta por muestreo	Censo	Encuesta basada en registros
Errores de cobertura	Errores de cobertura	Errores de relevancia
Errores de no respuesta	Errores de no respuesta	Errores de integración
Errores de medición	Errores de medición	
Errores de muestreo		

CAPÍTULO 11

Metadatos y sistemas informáticos

Es necesario documentar todas las encuestas, matrices de datos y bases de datos. De este trabajo de documentación se derivan *metadatos estadísticos*: información que describe los datos estadísticos y los procesos de la encuesta. Distinguimos entre *micro-metadatos*, que describen el contenido de las matrices de datos con microdatos (es decir, datos referidos a las unidades u objetos estadísticos individuales), y *macro-metadatos*, que describen el contenido de los tabulados estadísticos (es decir, datos referidos a los macrodatos que se han formado al agregar datos para los grupos de objetos). Sólo abordaremos los metadatos micro.

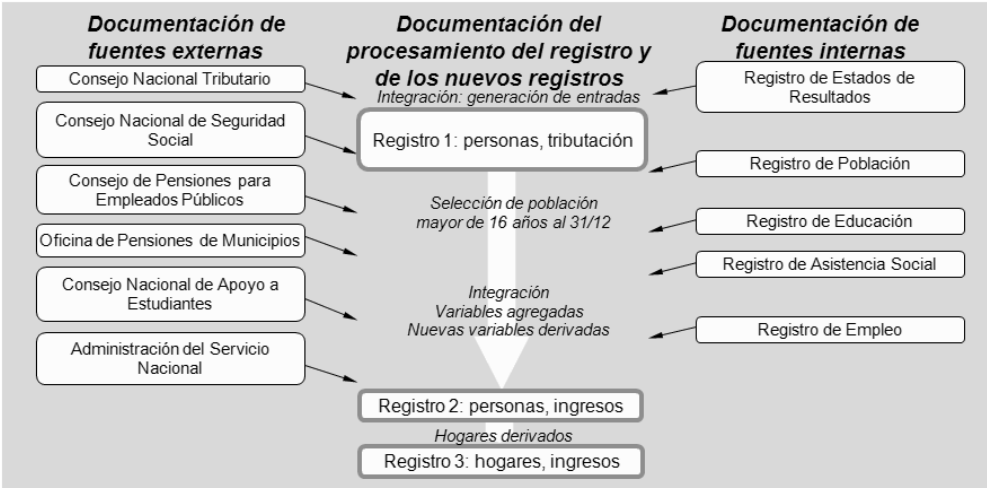
Los metadatos micro son necesarios tanto para los realizadores como para los usuarios de una encuesta. No obstante, sólo veremos las necesidades de metadatos de quienes crean registros estadísticos. En específico, más que las soluciones técnicas, analizaremos la necesidad de que el sistema de registros cuente con metadatos.

Luego de los metadatos, veremos los sistemas informáticos en general. Hasta ahora hemos hablado de los registros estadísticos en un nivel conceptual, es decir, cómo debe ser la presentación visual y el funcionamiento de los registros y las matrices de datos desde el punto de vista lógico. En este capítulo, continuaremos con ese razonamiento conceptual, pero también abordaremos la realización física cuando se crean bases de datos y aplicaciones.

11.1 REGISTROS PRIMARIOS: LA NECESIDAD DE METADATOS

Los registros estadísticos se crean integrando diversos registros fuente. Las estadísticas basadas en registros tienen requerimientos especiales en cuanto a metadatos, que difieren de las necesidades de metadatos de las encuestas con una captación de datos propia. El ejemplo del gráfico 11.1 muestra las fuentes necesarias para crear un registro y la necesidad de contar con distintos tipos de metadatos.

Gráfico 11.1 Registro de Ingresos y Tributación de la Oficina de Estadística de Suecia: la necesidad de metadatos



Para crear el Registro de Ingresos y Tributación sueco, se usan datos administrativos de seis autoridades distintas, junto con datos importados de cinco registros estadísticos suecos. Los microdatos constan de alrededor de 500 variables; para entender estas variables, se deben conocer bien las reglas fiscales que determinan su contenido. Debemos considerar que se pueden añadir nuevas variables porque el sistema tributario cambia constantemente y los nombres de las variables pueden cambiar en las fuentes administrativas.

Por lo anterior, cuando hablamos de la naturaleza de los metadatos, hay diferencias significativas entre las encuestas basadas en registros y las encuestas con una captación de datos propia.

También se debe distinguir entre la documentación de *registros* y la de *encuestas basadas en registros*. Cuando se usan registros que ya existen para crear nuevos registros, es crucial la documentación del *registro*. Este tipo de documentación se caracteriza por:

– el volumen de los metadatos, que puede ser muy alto;
– dar cuenta de cada una de las fuentes administrativas;
– dar cuenta de los cambios en el sistema administrativo;
– ser precisa, pues las variables pueden ser complicadas;
– dar cuenta de toda la parte del procesamiento de los registros destinada a crear objetos, variables y conjuntos de objetos.

Esto significa que el sistema de metadatos debe adaptarse para satisfacer los requisitos del sistema de registros y las encuestas basadas en registros.

11.1.1 Documentación de las fuentes administrativas

Los proveedores de datos presentan *descripciones de las entradas del registro*, etc., que indican la estructura y el contenido de los datos que entregan. Por su parte, el receptor debe conocer los *cuestionarios con las instrucciones* que se usaron para recopilar los datos administrativos. Los cuestionarios y las instrucciones se deben convertir a un formato electrónico para guardarlos en el sistema de metadatos y que todos los usuarios del registro tengan fácil acceso a ellos.

Los responsables de mantener contacto con los proveedores de datos también deben *entrevistarlos* para obtener más información de referencia. De igual modo, estas entrevistas se deberán documentar y guardar en el sistema de metadatos.

Es importante registrar cuidadosamente todos los cambios y guardarlos, de modo que sea fácil tener un panorama general de los datos y evaluar su comparabilidad a lo largo del tiempo. Por consiguiente, un sistema de metadatos también debe contener un *calendario*, que es un sistema informático con metadatos formalizados donde se puede buscar la información por *fecha*, *registro* y *variable*.

Los datos administrativos recibidos del proveedor se reestructuran dentro de la oficina de estadística y se someten a un primer proceso de validación para entonces crear una matriz. Los receptores del paquete de datos deben elaborar su propia documentación al respecto, incluido el procesamiento que se ha llevado a cabo.

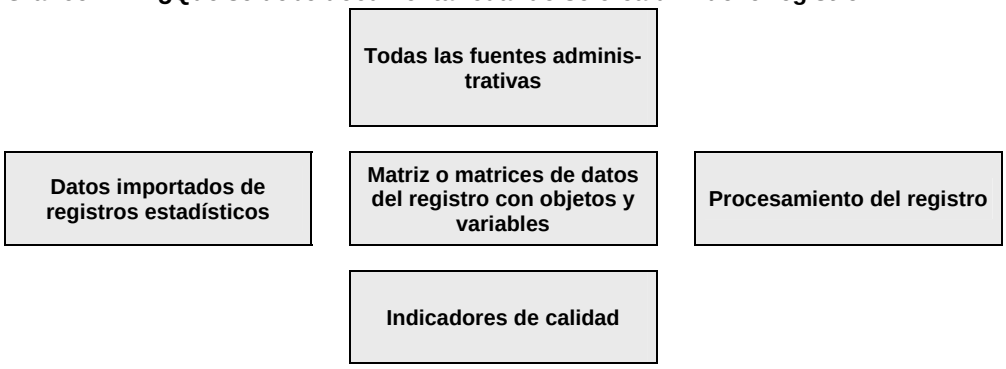
11.1.2 Documentación de las fuentes dentro del sistema de registros de la oficina de estadística

En la sección 3.3.5 se describen distintos tipos de variables que se deben documentar de manera diferenciada. Cuando se importan variables de otros registros estadísticos que forman parte del sistema de registros de la oficina de estadística, es fundamental tener fácil acceso a los metadatos disponibles para buscar y seleccionar las variables adecuadas. Al mismo tiempo que se importan los microdatos, los metadatos de las variables importadas podrían transferirse fácilmente a la documentación del nuevo registro. Esto evitaría la duplicación del trabajo, pues puede aprovecharse la documentación ya disponible. Para volver a usarla de manera apropiada, esta documentación deberá estar estrictamente formalizada de acuerdo con reglas comunes y almacenada en una base de datos de fácil acceso.

11.1.3 Documentación del nuevo registro

El nuevo registro creado se documenta usando la documentación de las fuentes administrativas mencionadas en la sección 11.1.1 e importando parte de la documentación de los diferentes registros que sean necesarios del sistema de la oficina de estadística. Entonces se puede complementar la documentación con descripciones del procesamiento y la integración realizados, así como de las variables derivadas que se generaron.

Gráfico 11.2 ¿Qué se debe documentar cuando se crea un nuevo registro?



11.2 CAMBIOS A LO LARGO DEL TIEMPO: NECESIDAD DE METADATOS

Hay cuatro tipos de eventos que pueden afectar las estadísticas basadas en registros, de modo que, para evitar interpretaciones incorrectas de las series de tiempo de las encuestas basadas en registros, necesitamos saber lo siguiente:

– ¿Ha habido cambios en el sistema administrativo del que provienen las fuentes, de modo que se ha dado una nueva definición a conceptos administrativos?

– ¿Ha habido cambios en la manera en que la oficina de estadística crea el registro? Por ejemplo, ¿hay acceso a nuevas fuentes o se han introducido nuevos métodos de estimación? Se debe dar seguimiento a varios indicadores de calidad para no malinterpretar los datos.

– ¿Ha habido cambios en las clasificaciones usadas en el registro? Por ejemplo, tal vez han cambiado los códigos de clasificación industrial o nivel de escolaridad.

– ¿Han ocurrido cambios externos que podrían haber afectado indirectamente las estadísticas? Si, por ejemplo, cambia la tasa del impuesto al valor agregado (IVA), no sólo se ven afectados los datos sobre el IVA, sino todos los datos relacionados con el consumo privado. Para interpretar correctamente las estadísticas, debemos conocer estos cambios.

Por lo tanto, se debe elaborar un calendario de eventos para documentar y compilar en un solo sitio los eventos que podrían afectar los diferentes registros. Esto facilitará las cosas para los generadores de registros estadísticos y los usuarios de las encuestas basadas en registros, y, en consecuencia, disminuirá el riesgo de una interpretación incorrecta de los datos.

Ese calendario puede contener información básica y detalles sucintos sobre:

– qué ocurrió;

– cuándo ocurrió;

– qué registro o registros y qué variables resultaron afectados;

– qué efecto tuvo el evento;

– referencias o fuentes de información adicional.

Gráfico 11.3 Ejemplo de contenido de un calendario de eventos

Evento:	Introducción de la Iniciativa de Educación para Adultos
Fecha:	1997
Registro:	Municipal de Educación de los Adultos, año 1997
Variable:	Nuevas variables agregadas (mostradas, un elemento por variable)
Efecto:	Aumento del número de estudiantes en aproximadamente 40 por ciento
Evento:	Nuevas rutinas dan información con mayor rapidez sobre emigrantes que han cesado sus actividades empresariales
Fecha:	1997
Registro:	Registro Empresarial
Variable:	-
Efecto:	Mayor confiabilidad dada la menor sobrecobertura
Evento:	Introducción de la nueva Clasificación Industrial Sueca, SNI 92
Fecha:	1995
Registro:	Registro de Empleo 1993 (elementos correspondientes para todos los registros relacionados)
Variable:	Rama industrial
Efecto:	El nuevo SNI 92 sustituye la antigua Clasificación Industrial Sueca, SNI 69
Evento:	Checoslovaquia se divide en dos nuevos estados: República Checa y Eslovaquia
Fecha:	01-01-1993
Registro:	Registro de Población
Variable:	Ciudadanía
Efecto:	-

La información sobre *registro* y *variable* en el calendario de eventos puede servir como vínculos con la información en otras bases de datos con metadatos formalizados.

11.3 REGISTROS INTEGRADOS: NECESIDAD DE METADATOS

Si, por ejemplo, se crea un registro integrado longitudinal con *diez* versiones anuales de tres registros estadísticos diferentes, será necesario usar eficazmente los metadatos disponibles. Supongamos que los *tres* registros fuente para cada año contienen *200* variables y el sistema de metadatos contiene *30* definiciones para las diferentes poblaciones de registro y *2 000* definiciones de variables. Esto ilustra que el proceso del registro puede implicar grandes cantidades de metadatos.

Si, por ejemplo, los registros 1 y 2 tienen definiciones de población de registro que no han cambiado en diez años y el registro 3 una población de registro cuya definición se modificó una vez, sólo se necesitan 4 (= 1 + 1 + 2) de las 30 posibles definiciones de población.

Si, por ejemplo, cada año se ha importado un total de 50 variables de los tres registros al nuevo registro integrado, pero sólo ha cambiado la definición de cuatro variables, una vez cada una, durante el periodo de diez años, entonces sólo se necesitan 54 (= 50 + 4) de las 2 000 definiciones de variables.

Este ejemplo nos muestra la necesidad de contar con un sistema de metadatos eficiente sin grandes cantidades de metadatos redundantes. En este caso, las cuatro definiciones de población y las 54 definiciones de variables deberán ser de fácil acceso.

11.4 BASE DE DATOS DE CLASIFICACIONES Y DEFINICIONES

En el caso de las estadísticas de registros, que se basan en datos administrativos, reviste particular importancia poder estudiar las definiciones de las variables y compararlas a lo largo del tiempo. El Registro de Ingresos y Tributación del gráfico 11.1 ilustra esta necesidad: se deben manejar alrededor de 500 variables basadas en reglas administrativas. Muchas de esas reglas y variables se modifican cada año.

11.4.1 Base de datos de clasificaciones

Las clasificaciones industrial y de ocupación, así como los catálogos de productos, nivel de escolaridad y de regiones son ejemplos de *estándares* y *clasificaciones* estadísticos importantes. Las fuentes administrativas contienen datos sobre estas clasificaciones ordenadas jerárquicamente, información que se aprovecha para crear variables dentro del sistema de registros. Estas clasificaciones se modifican a intervalos regulares. Como los conjuntos de valores (conjuntos de todos los códigos o categorías) también son grandes, se requiere una base de datos de clasificaciones para manejar todos los códigos y claves entre las distintas versiones. Esta base de datos es un valioso recurso cuando se documentan las variables de un registro.

11.4.2 Base de datos de definiciones y variables derivadas

Así como son necesarias herramientas informáticas para manejar la definición de las clasificaciones estadísticas, se requiere una herramienta con metadatos formalizados para manejar la gran cantidad de definiciones de variables complejas que van cambiando con el paso

del tiempo. Ilustramos esto con un ejemplo ficticio en el que se han documentado tres años de un registro de ingresos (RI) y también hay un registro de ingresos longitudinal (RIL) que integra los tres años.

Gráfico 11.4 Documentación de las variables de registro mediante una base de datos de definiciones

Registro	Nombre de la variable	Código de definición
RI 2001	Subsidio por enfermedad	SEn1
	Subsidio por embarazo	SEm1
	Pago de incapacidad por enfermedad	IEn1
RI 2002	Subsidio por enfermedad	SEn2
	Subsidio por embarazo	SEm1
	Pago de incapacidad por enfermedad	IEn2
RI 2003	Subsidio por enfermedad	SEn2
	Subsidio por embarazo	SEm2
	Pago de incapacidad por enfermedad	IEn3
RIL	Pago de incapacidad por enfermedad 2001	IEn 1
	Pago de incapacidad por enfermedad 2002	IEn 2
	Pago de incapacidad por enfermedad 2003	IEn 3

Base de datos de definiciones

Código	Definición	Definición usada	
		Primera vez	Última vez
SEn1	SEn1 = "……"	2001	2001
SEn2	SEn2 = "……"	2002	-
SEm1	SEm1 = "……"	2001	2002
SEm2	SEm2 = "……"	2003	-
IEn1	IEn 1 = SEn1 + SEm1	2001	2001
IEn2	IEn 2 = SEn2 + SEm1	2002	2002
IEn3	IEn 3 = SEn2 + SEm2	2003	-

El registro de ingresos ficticio contiene tres variables de ingreso:

– *Subsidio por enfermedad,* sujeto a nuevas reglas desde 2002.

– *Subsidio por embarazo,* sujeto a nuevas reglas desde 2003.

– *Pago de incapacidad por enfermedad,* una variable derivada, el total de los subsidios por enfermedad y embarazo.

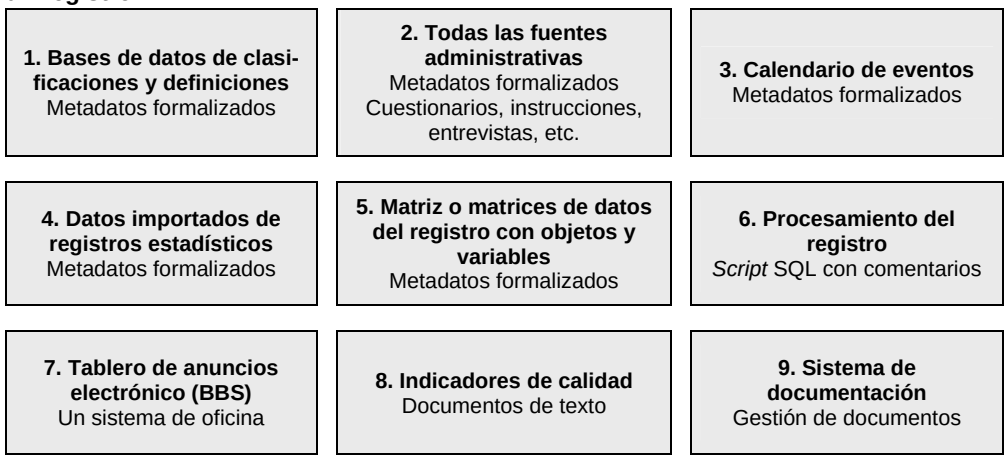
El registro longitudinal sólo contiene la variable *Pago de incapacidad por enfermedad* para cada año. El gráfico anterior muestra que hay variables con el mismo nombre, como *Subsidio por enfermedad,* que pueden tener definiciones distintas. Además, hay variables con nombres diferentes, como *Pago de incapacidad por enfermedad* en el RI 2001 y *Pago de incapacidad por enfermedad 2001* en el RIL, que pueden tener la misma definición.

Sin embargo, como los códigos de definición son únicos (es decir, se usa un código específico en todo el sistema de registros para una y sólo una definición de variable), no debe haber interpretaciones erróneas. También es más fácil dar seguimiento a los cambios de definiciones con una base de datos de definiciones.

11.5 NECESIDAD DE METADATOS PARA REGISTROS

Los generadores de registros estadísticos dentro del sistema de registros requieren diferentes tipos de metadatos y herramientas informáticas prácticas para compilar y aprovechar estos metadatos. A continuación mostramos nueve tipos de metadatos y las herramientas que se podrían utilizar.

Gráfico 11.5 Diferentes tipos de metadatos y herramientas en la documentación de un registro



Debe haber un sistema que integre todos los metadatos formalizados que existen actualmente, pero también los que hay en el calendario y las bases de datos de clasificaciones y definiciones que mencionamos antes. Además se requiere un sistema de gestión de documentos con otros metadatos. Los sistemas con *metadatos formalizados* se pueden usar para lo siguiente (los números se refieren al gráfico 11.5):

1. Bases de datos de clasificaciones y definiciones de fácil acceso.
2. Documentación de matrices de datos procedentes de fuentes administrativas.
3. Calendario de eventos, fácil acceso a la información sobre cambios importantes.
4. Datos importados de los registros estadísticos, los metadatos formalizados se importan fácilmente.
5. Documentación de las matrices de datos en el registro estadístico. Se describe la población de registro, el tipo de objetos y las variables.

El resto de la documentación puede consistir en *diferentes tipos de documentos* (los números se refieren al gráfico 11.5):

2. La información en texto sobre los sistemas administrativos, los cuestionarios administrativos con instrucciones y las minutas y notas de reuniones con los proveedores de registros.
6. *Script* SQL con comentarios que describen el procesamiento del registro.
7. Un tablero de anuncios electrónico (BBS) para los usuarios de registros que detectan inconsistencias y errores. De acuerdo con la sección 5.4.7, los responsables de los registros base deben incluir sus contribuciones en el tablero de anuncios correspondiente a su registro base.
8. Los indicadores de calidad son los indicadores más importantes para el registro en cuestión.
9. Todos los documentos anteriores se gestionan por medio de un sistema especial que facilita su acceso.

Documentos de texto de formato uniforme

Normalmente las matrices de datos creadas mediante la captación de datos en encuestas por muestreo se documentan en textos estructurados de manera uniforme. En el gráfico

siguiente se ilustra cómo adecuar este tipo de documentación a las encuestas basadas en registros. Se compara la parte más importante de cada tipo de encuesta: el proceso de captación de datos para las encuestas por muestreo y el proceso de integración para las encuestas basadas en registros.

Gráfico 11.6 Metadatos para encuestas por muestreo y encuestas basadas en registros

Encuesta por muestreo: la matriz de datos **Proceso de captación de datos**	**Encuesta basada en registros: el registro** **Proceso de integración**
1. Marco y procedimiento del marco 2. Procedimiento de muestreo 3. Cuestionario 4. Procedimiento de captación de datos 5. Preparación de datos	1. Descripción de fuentes 2. Recepción y validación de cada fuente 3. Integración 1: población de registro 4. Integración 2: objetos 5. Integración 3: variables 6. Validación de consistencia

En este gráfico vemos la manera en que se crean los microdatos, que es distinta para estos dos tipos de encuestas. Por lo tanto, esta parte de la documentación debe diferir. Las otras partes de la documentación pueden tener la misma estructura.

El sistema de metadatos: encuesta con captación de datos

Una oficina de estadística recolecta metadatos por medio de sistemas especiales a cargo de su personal. Los responsables de la documentación llenan formularios electrónicos que deben dar como resultado metadatos de buena calidad. Buena calidad significa que el sistema de metadatos tiene una buena cobertura, un bajo índice de no respuesta y errores de medición pequeños, así como metadatos fáciles de consultar y entender.

Definir los tipos de objetos puede ser una parte difícil de la documentación, pues pueden surgir errores de medición o clasificaciones incorrectas si los conceptos son difíciles de entender. La distinción entre *tipo de objeto* y *variable* debe quedar clara para quienes se ocupan de documentar los metadatos; de lo contrario, los usuarios de los metadatos tendrán problemas cuando busquen datos sobre un tipo de objeto en particular. Si un tipo de objeto se ha definido como otro tipo o como una variable, el usuario no encontrará los metadatos que busca. Para evitar estos errores de clasificación o medición, sólo se deben definir unos cuantos tipos de objeto en el sistema, que deberán ser de fácil comprensión.

Ejemplo: Un niño en edad escolar se puede definir como un objeto relacional *persona • escuela* o como una *persona*. Recomendamos definir a los niños en edad escolar como personas; será sencillo encontrar otros registros sobre personas con más variables concernientes a niños en edad escolar.

Ejemplo: Los productos se pueden definir como *variables* vinculadas con empresas o como *tipo de objeto*. Recomendamos que los productos se definan como variables: las empresas elaboran productos de distintos tipos, entonces el valor y la cantidad de cada producto se pueden definir como variables de empresas que se pueden combinar con otras variables de empresas de otros registros.

En un entorno informático, el término “objeto” es de uso frecuente y en ocasiones se le define de manera distinta al término usado en estadística. En una herramienta para el manejo de bases de datos, los renglones de ciertas tablas de bases de datos reciben el nombre de “objetos” sin que sean objetos en el sentido estadístico o conceptual. Esto puede dar pie a malentendidos. Cuando se documenta una encuesta, sólo los objetos que forman parte de la población de registro deben llamarse “objetos” en la parte estadística de la documentación.

11.6 SISTEMAS INFORMÁTICOS PARA ESTADÍSTICAS BASADAS EN REGISTROS

Cuando se diseña un sistema informático como soporte para la creación de registros estadísticos, la documentación de los registros es un factor crucial. Por lo tanto, resulta esencial que el trabajo de documentación se haya iniciado antes del diseño de sistemas. Los diversos tipos de metadatos formalizados que mencionamos previamente estarán disponibles y podrían usarse con alguna de las herramientas informáticas que existen para el manejo de los datos.

En esta sección, abordaremos el análisis, el diseño y la realización física de los sistemas informáticos.

11.6.1 Análisis

El proceso de análisis tiene por objeto sentar las bases para el diseño y la realización del sistema mediante el análisis y la formalización de los requisitos que debe satisfacer. El método aplicado consiste en modelar —identificar, examinar y documentar— los objetos y procesos del sistema.

En la modelación de objetos se identifican los datos de entrada y de salida usados en el sistema y se describen las relaciones entre ellos.

La modelación de procesos ilustra los procesos del sistema y describe tanto su secuencia como las relaciones de dependencia entre ellos. En esta etapa se describe cómo se captan los datos. También se documenta la manera en que los usuarios deberían, o les gustaría trabajar, mediante la descripción de *casos de uso* de una manera estandarizada.

Un sistema informático para estadísticas basadas en registros difiere de uno para encuestas con captación de datos propia, sobre todo en lo relativo a la modelación del proceso de captación de datos. En vez de definir una población marco y seleccionar una muestra, el proceso incluye determinar en qué medida los registros que componen las fuentes de datos principales satisfacen la necesidad de datos de salida. Si se requieren más datos, se deben identificar y documentar fuentes adicionales. En esta etapa, la principal tarea consiste en modelar los procesos de selección y emparejamiento necesarios para crear el registro.

En esta etapa también se debe investigar cómo funcionará la cooperación con otros sistemas. La principal cooperación que se debe explicar es dentro del sistema de registros: qué relaciones de dependencia existen respecto a otros registros, cómo manejar los conflictos entre fuentes de información, etc. También se debe explicar el papel del registro como fuente para otros registros dentro del sistema de registros.

El manejo de la *dimensión temporal* (véase sección 3.2.4) es una complicación típica de los registros estadísticos. Una encuesta periódica con captación de datos propia creará, por decirlo de manera simple, un archivo o una tabla de base de datos nuevos en cada ronda de la encuesta, o agregará nuevas entradas a un archivo o una tabla de base de datos que ya existe. Por su parte, un registro estadístico, si lo permiten los registros subyacentes, estará en constante cambio mediante la adición, eliminación o modificación de datos. En el proceso de análisis se debe definir claramente qué información de referencia sobre el tiempo será necesaria o se esperará para la producción regular, así como las tareas específicas previstas. Así se establecerán los criterios para decidir qué tipo de información es necesario en el registro de que se trate: la población actual, poblaciones en momentos determinados y arbitrarios, población por año calendario, eventos e información longitudinal.

El proceso de análisis también debe mostrar el nivel de detalle que se requiere para la información utilizada en diferentes partes del sistema. Asimismo, se debe documentar la frecuencia posible, deseable y significativa para la actualización el registro estadístico desde el registro subyacente.

Cabe destacar la importancia de mantener una visión de conjunto del sistema de registros durante la fase de análisis. Aunque el trabajo de desarrollo se concentre en un registro específico, constantemente se debe mirar más allá y tomar nota de los requerimientos de todo el sistema de registros.

11.6.2 Diseño

Considerando lo que se espera del sistema, en el proceso de diseño se deben tomar decisiones sobre los siguientes aspectos:

- la selección de tecnología para la solución total de sistemas
- cualquier división en subsistemas
- el diseño de la base de datos
- una arquitectura para cada aplicación o procesamiento

Selección de tecnología

Definir la cooperación entre un registro estadístico y sus registros fuente es una parte importante del diseño del sistema. Cuando se diseña un registro primario, la cooperación técnica con el registro administrativo es de particular importancia. Esto dependerá de varios factores, entre ellos:

- La tecnología de almacenamiento usada por el registro administrativo: ¿los datos se pueden transferir directamente o se deben convertir?
- Frecuencia de actualización: Si los datos sólo se transferirán en raras ocasiones, quizás una vez al año, se podría aplicar una solución manual relativamente simple. Sin embargo, una mayor frecuencia de actualización requerirá un mayor nivel de automatización y exactitud tanto del emisor como del receptor.
- Cantidades de datos: La transferencia de una gran cantidad de datos naturalmente impondrá mayores exigencias a la solución técnica. Se puede transferir una pequeña cantidad por correo electrónico, CD-ROM o un dispositivo similar, en tanto que cantidades más grandes requerirán soluciones bien planeadas.
- Seguridad/acceso no permitido: Todos los microdatos relacionados con objetos individuales son estrictamente confidenciales y se deben manejar de acuerdo con la reglamentación vigente en materia de confidencialidad.
- Seguridad/vulnerabilidad: Se refiere a los riesgos de perder datos cuando se transfieren mediante, por ejemplo, rutinas o tecnología inadecuadas, y también a riesgos relacionados con la pérdida de competencia de ambas partes.
- Tecnología de comunicación: Normalmente hay varias tecnologías disponibles para todos los casos. Al elegir una tecnología se deben tener en cuenta los factores antes mencionados.

Además de que el registro debe ser capaz de cooperar con uno o más registros fuente, los aspectos técnicos también deben funcionar con otros registros del sistema. Por consiguiente, la tecnología y las soluciones aplicadas anteriormente a registros con una vinculación estrecha tendrán una gran influencia en las soluciones para el registro de que se trate.

El análisis y la evaluación combinada de estos requerimientos, características deseables y posibilidades pueden ocasionar diferencias considerables en el trabajo que se realice en las siguientes etapas de diseño.

Diseño de las bases de datos

El diseño de las bases de datos es crucial para el funcionamiento eficaz de un sistema. Por lo tanto, habrá de ser responsabilidad de programadores de sistemas con una amplia experiencia en modelación de datos.

El diseño de una base de datos para un sistema informático que trabajará con estadísticas basadas en registros no difiere en lo fundamental del diseño de bases de datos para otros sistemas informáticos destinados a la producción de estadísticas. El sistema se debe optimizar para obtener:

– facilidad de uso;
– confiabilidad y buen desempeño para cargar datos y valores nuevos y actualizados,
– un procesamiento eficiente, es decir, validación, derivación de variables, etcétera;
– datos de salida, es decir, producción de tablas estadísticas y otras tareas normales o especiales.

Muchas veces esto supone requerimientos discrepantes para el diseño de bases de datos y lleva a dividir el sistema en varias bases de datos. Con frecuencia se mencionan tres tipos principales de bases de datos optimizadas para diferentes aplicaciones: bases de datos de entrada, de procesamiento y de salida.

– Una *base de datos de entrada* debe tener capacidad para recibir todo tipo de actualizaciones, incluidos datos de entrada totalmente incorrectos. Por consiguiente, las reglas de consistencia de estas bases suelen ser mínimas. Si es necesario, los datos se verifican y corrigen antes de enviarlos.
– Una *base de datos de procesamiento* se normaliza a menudo, es decir, está sujeta a estrictas reglas de consistencia y no tiene datos duplicados. Se trata de un modelo optimizado para la edición de microdatos con riesgo mínimo en caso de que se introduzcan datos inconsistentes.
– Una *base de datos de salida* se debe optimizar para los tipos más frecuentes de selección, análisis y otros procesamientos. Una vez cargados los datos, no se deben hacer cambios. Así pues, en lo que respecta al contenido, la base de datos de salida es equivalente de la matriz de datos completa. Se permiten los duplicados en caso de que mejoren el desempeño o faciliten las cosas a los usuarios.

El formato de datos para todas las variables se debe determinar en esta fase. La mayoría de las veces esto no causa problemas, pero en el caso de las variables de vinculación, en particular las que pueden usarse para vincular diferentes registros, es muy importante apegarse a normas establecidas, por ejemplo, los números de identificación personal (NIP) suecos se deben guardar usando exactamente 12 dígitos.

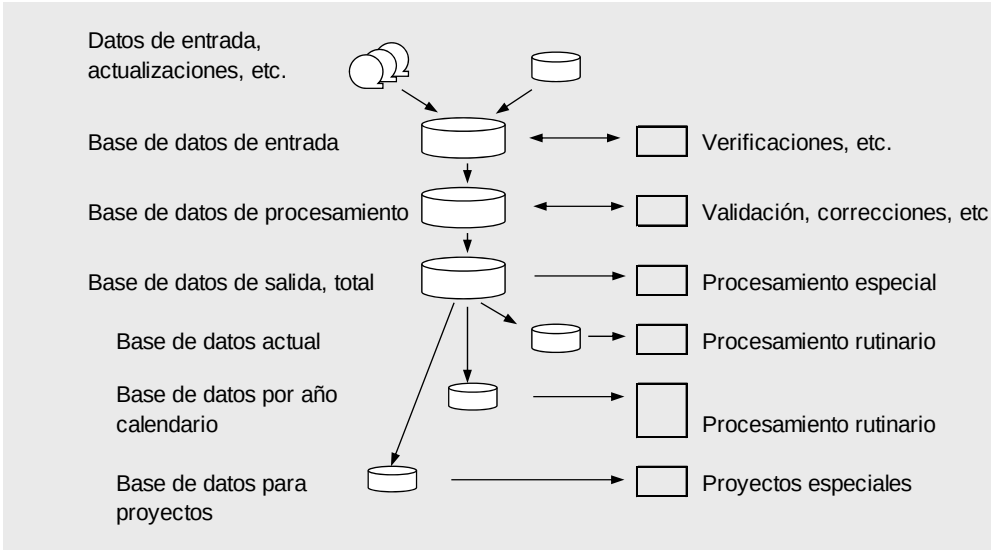
La dimensión temporal (véase sección 11.6.1) puede ser un factor de mucho peso en el diseño de las bases de datos para un registro base. Si el sistema tiene grandes requerimientos de información sobre tiempo para cada modificación y abarca un periodo largo, existe el riesgo de que usar una sola base de datos de salida para todo el procesamiento de salida resulte un sistema complicado para el usuario y con un bajo nivel de desempeño. En esos casos, tal vez sea necesario dividir más la base de datos de salida, por ejemplo:

– una base de datos total o histórica;

– una o varias bases de datos para momentos específicos, como final de año, y una base de datos para la versión de año calendario del registro;
– una base datos para la versión de inventario actual del registro.

Del mismo modo, en el caso de un trabajo especial de gran alcance basado en el registro puede ser necesario crear bases de datos independientes para ese proyecto. Llevar a cabo las tareas especiales y el procesamiento rutinario en las mismas bases de datos puede ocasionar serias perturbaciones, que se deben evitar.

Gráfico 11.7 Diferentes tipos de bases de datos para un registro estadístico



El requisito de la cooperación dentro del registro obviamente impone ciertos límites a la selección del sistema de gestión de bases de datos. Un programa de cómputo que, en ciertos aspectos, tiene un excelente desempeño o funciona bien para determinado registro, tal vez no se elija si no funciona bien —o no funciona en absoluto— con el programa que utilizan registros muy relacionados dentro del sistema de registros.

Modelación para el sistema de registros, no sólo el registro

A lo largo de este libro, hemos destacado la importancia de tener una visión de conjunto, el sistema de registros, y no sólo de cada registro. Podemos ir más allá y afirmar que el sistema de registros es un componente fundamental del *almacén de datos estadísticos*.

El término *almacén de datos* se ha usado sobre todo en el mundo empresarial para describir un repositorio coordinado de información, consistente y de calidad asegurada, a menudo relacionado con análisis longitudinales de datos principalmente financieros.

Un almacén de datos (Data Warehouse) suele contener grandes cantidades de datos, divididos en un gran número de categorías de información. Cuando se diseña una o varias bases de datos para un almacén de datos, es preferible minimizar la necesidad de emparejar más de dos *tablas de bases de datos* al mismo tiempo, en parte porque desde el punto de vista lógico y sintáctico el emparejamiento es difícil para los usuarios y en parte porque deteriora el desempeño del sistema de gestión de bases de datos.

Variables derivadas y vistas

Siempre que el diseño del sistema sea conceptual, o lógico, las variables derivadas se manejan de la misma forma que otras variables. Una parte importante del diseño de la base de datos física consiste en determinar qué variables derivadas se deben almacenar permanentemente y cuáles se deben combinar y calcular sobre demanda. Esta última alternativa puede ser por circunstancias específicas o para una mayor facilidad de uso creando *vistas*. Una *vista* es una "tabla de base de datos virtual", es decir, una expresión predefinida que a ojos del usuario es como cualquier tabla de base de datos, pero que en realidad lleva a cabo operaciones en una o varias de las tablas de base de datos subyacentes.

Ejemplo: La tabla de base de datos X contiene datos sobre diversos tipos de ingreso, horarios de trabajo, etc. A partir de los datos contenidos en X, podemos derivar la variable *salario de tiempo completo*, que colocamos en la vista Y. Entonces los usuarios pueden hacer una simple consulta para seleccionar la información sobre salario de tiempo completo de Y, aunque cada vez el cálculo se llevará a cabo en X. La ventaja con una *vista* es que resulta fácil de entender y usar. Una desventaja es que para funcionar bien puede requerir un equipo de cómputo de alto desempeño.

Para decidir si una variable derivada se almacenará físicamente en una tabla de base de datos o virtualmente en una *vista*, se debe considerar el efecto que esto puede tener en el desempeño, qué riesgos puede plantear un almacenamiento duplicado y con qué frecuencia se usará la variable.

El desempeño de una vista depende de la cantidad de datos (número de objetos), el número de tablas de base de datos afectadas y la complejidad de la expresión subyacente. Si es posible derivar una variable por medio de un cálculo simple a partir de unas cuantas variables presentes en una o dos tablas dentro de la misma base de datos, entonces el uso de una *vista* en vez de almacenar la nueva variable en una tabla de base de datos normalmente no afectará de manera perceptible el desempeño.

No obstante, el desempeño sí puede verse considerablemente mermado si la variable se deriva de otra base de datos, quizás un servidor remoto, y se procesan grandes cantidades de datos. Si se trata de una variable que se usará con frecuencia, tal vez sea necesario almacenarla de manera local.

En la mayoría de los contextos, las *vistas* se deben considerar de la misma manera que cualquier otra tabla de base de datos y, por consiguiente, se debe documentar de la misma manera. Las *vistas* creadas sólo con el propósito de preparar un emparejamiento recurrente entre tablas de bases de datos pueden ser una excepción.

Arquitectura de la aplicación

En esta fase, también se debe investigar la necesidad de diseñar aplicaciones a la medida. Un registro con actualizaciones muy frecuentes de los datos de entrada requeriría un proceso automatizado para la transferencia y carga, mientras que para un registro con actualizaciones irregulares o anuales, por ejemplo, basta con rutinas manuales. De igual forma, un procesamiento rutinario estandarizado y muchas veces recurrente puede simplificarse diseñando una aplicación que "cualquiera" pueda manejar en vez de repetir secuencias de manipulación similares para cada tarea.

11.6.3 Realización física

La realización física se refiere sobre todo a la programación y establecimiento de las bases de datos necesarias a fin de preparar el sistema para su uso. La programación y la creación

de bases de datos se debe llevar a cabo con las herramientas recomendadas y apegándose a las normas establecidas.

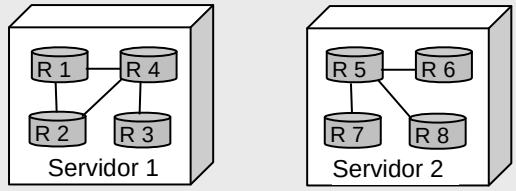
Optimización física del sistema de registros

Las bases de datos que integran el sistema de registros deben modelarse, de modo que se establezca una buena cooperación lógica entre ellas, y se obtenga un buen funcionamiento físico en conjunto. No es posible establecer un sistema de registros en una sola base de datos física, probablemente ni siquiera en un solo servidor. La cooperación física estará afectada por el hecho de que el emparejamiento se haga en un servidor o entre diferentes servidores.

Las herramientas para permitir que los usuarios interpreten un grupo de servidores como una unidad lógica están en constante mejoramiento, pero lo cierto es que un servidor será mucho más veloz internamente que cualquier enlace de comunicación entre servidores. Se debe valorar cuáles son los registros dentro del sistema con mayor necesidad de funcionar conjuntamente, para entonces ubicarlos en el mismo servidor.

En el ejemplo del gráfico 11.8, la información de los registros R1–R4 regularmente se usa de manera conjunta, como ocurre con la información de los registros R5–R8, de modo que las bases de datos se agrupan de ese modo en los dos servidores.

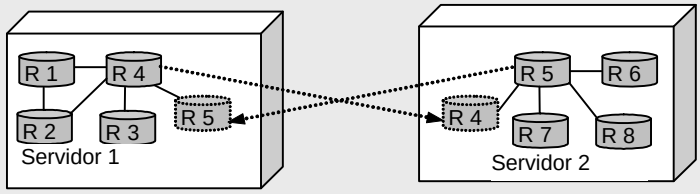
Gráfico 11.8 Los registros que se usan conjuntamente se ubican en el mismo servidor



Una ubicación de registros como la que se muestra en el ejemplo no impide usar conjuntamente los registros R4 y R5, por ejemplo. Sólo será un poco más complejo y puede llevar más tiempo. Un método para retener el uso en los dos servidores del ejemplo, y mejorar el uso entre los registros R4 y R5, consiste en mantener copias de estos registros en el otro servidor.

Un almacenamiento doble controlado con esta configuración se puede manejar de diversas maneras. Probablemente la más segura sea usar la tecnología de replicación del sistema de gestión de bases de datos. En el ejemplo del gráfico 11.9, cada uno de los cambios en el registro R4 del servidor 1 también se hará en el registro 4 que está en el servidor 2 en determinado momento, por ejemplo, cada hora, cada noche o cada semana. Lo mismo se aplica al registro 5 del servidor 2 y su réplica en el servidor 1.

Gráfico 11.9 Almacenamiento duplicado de algunos registros usados conjuntamente



Indexación

Como todo el procesamiento de los datos de un almacén de datos se apoya en bases de datos relacionales, es fundamental una *indexación* correcta y apropiada para el buen desempeño. Sin índices apropiados, el sistema de gestión de bases de datos siempre evaluará cada combinación posible antes de tomar una decisión cuando, por ejemplo, empareje dos tablas de bases de datos, aunque el usuario sólo quiera emparejar un renglón de cada tabla. En cambio, al indexar las claves primarias y externas se dará al sistema la información necesaria para ubicar directamente los dos elementos que emparejará.

Por lo general, las herramientas de modelación de datos indexan automáticamente las tablas de bases de datos de acuerdo con las claves primarias y externas. En una base de datos de procesamiento, a menudo éstos son los únicos índices necesarios a fin de crear un buen entorno para la validación de datos. Indexar otras variables podría incluso deteriorar el desempeño.

Sin embargo, en una base de datos de salida, donde los cambios son inusuales y bien planeados, y las búsquedas y emparejamientos tal vez deban llevarse a cabo en variables distintas de las claves, una indexación bien planeada puede redundar en mejoras de desempeño significativas. Esto se logra a costa de un desempeño menos eficaz en la actualización.

11.6.4 Uso de registros

El uso de un sistema basado en registros para la producción de tablas estadísticas u otro procesamiento no tiene requisitos específicos en comparación con otros sistemas, excepto tal vez en dos aspectos. El sistema basado en registros con frecuencia contiene cantidades muy grandes de datos y supone la integración de varias fuentes de datos.

Grandes cantidades de datos

El procesamiento o la recuperación de grandes cantidades de datos ameritan ciertas consideraciones. El tiempo necesario para el procesamiento muchas veces mantiene una dependencia no lineal de la cantidad de datos. El tiempo suele depender del equipo y los programas de cómputo, la configuración y el método de programación. En casos difíciles, las limitaciones de la capacidad de memoria pueden originar una relación exponencial entre la cantidad de datos y el tiempo de procesamiento. Sin embargo, a menudo el factor más importante para el buen desempeño es el método de trabajo del usuario. Con cantidades de datos modestas, el usuario puede recurrir al ensayo y error apoyándose en corridas de prueba, lo que resulta imposible con grandes cantidades de datos.

Emparejamiento

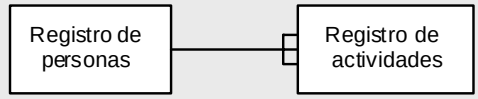
Emparejar información de varios registros o entre un registro y una encuesta por muestreo es una tarea común. En un nivel conceptual, parece bastante fácil, siempre que se disponga de las variables de vinculación necesarias.

No obstante, a la larga, cuando los registros conceptuales se conviertan en tablas de bases de datos, el emparejamiento (conocido como *unión —join—* en la terminología de SQL) se vuelve un problema tangible. El siguiente es uno de los ejemplos de la sección 3.3.5:

Un registro sobre personas que utiliza el número de identificación personal como clave primaria se puede emparejar con un registro de actividades en el que las actividades remuneradas se identifican con tres variables: número de identificación personal, número de establecimiento y número de organización. El número de identificación personal en el Registro de Actividades es la clave externa para el emparejamiento.

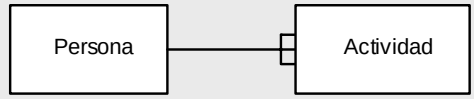
Una persona (un número de identificación personal) puede ocurrir sólo en un lugar en el registro de personas, pero en uno o varios lugares en el registro de actividades. Por lo tanto, tenemos una relación uno a muchos entre objetos. Esta situación se puede representar gráficamente de la siguiente manera:

Gráfico 11.10 Emparejamiento de dos registros



Cuando esto se realiza físicamente, en primer lugar podemos dar por sentado que los datos que queremos recopilar existen en una sola tabla de base de datos en el registro de personas (*Persona*) y en una sola tabla del registro de actividades (*Actividad*). Hasta ahora, sigue pareciendo bastante fácil emparejar ambas *tablas de bases de datos*:

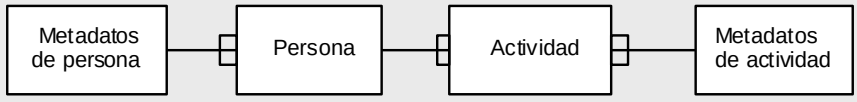
Gráfico 11.11 Relación persona–actividad



El emparejamiento se puede expresar en lenguaje de SQL como *unión* (*join*). Varias herramientas para usuarios finales, como SAS y SPSS, permiten efectuar una unión sencilla sin necesidad de que el usuario escriba el código SQL, siempre y cuando se hayan especificado correctamente las claves al crear las tablas de bases de datos. En la mayoría de los casos, es preferible usar estas herramientas a escribir el código SQL, pues el apoyo que ofrecen las herramientas reduce los riesgos de vínculos perdidos y errores de captura, entre otros.

No obstante, estas tablas de bases de datos normalmente sólo contienen los códigos numéricos de las variables cualitativas y en muchos casos queremos incluir el texto explicativo real guardado en las tablas de metadatos. De inmediato se complica el gráfico de nuestro sencillo ejemplo inicial, pues requiere una unión entre cuatro tablas de bases de datos:

Gráfico 11.12 Emparejamiento de dos registros con metadatos



Esto probablemente se complique más si, por ejemplo, el almacenamiento de las variables de actividades de nuestro interés está disgregado en varias tablas de bases de datos o se agregan algunas variables mientras se están emparejando.

Otra complicación frecuente es el manejo del *no emparejamiento* (véase sección 3.3.5). En nuestro ejemplo, puede ocurrir que en el registro de actividades haya objetos perdidos por algún motivo o que tengan un número de identificación personal incorrecto, y es casi un hecho que algunos de los objetos del registro de personas no tendrán un objeto correspondiente en el registro de actividades.

En SQL, como con todas las herramientas para usuarios finales (SAS, SPSS, etc.), la situación habitual es que sólo se seleccionan emparejamientos completos. Si queremos seleccionar todos los renglones de la tabla de base de datos *Persona* y añadir datos de *Actividad* cuando existen —o dejarlos en blanco si no existen (*valor nulo*) — esto se debe indicar específicamente usando la función *unión exterior* (*outer join*) de SQL. De manera sintáctica

y lógica, el uso de *unión exterior* es limitado, por lo tanto, se debe combinar con agregación y con emparejamientos completos (*unión interior* o *inner join*).

Gráfico 11.13 Unión interna (*inner join*) y unión externa (*outer join*)

Unión interna (*inner join*): Sólo selecciona los renglones donde el número en *Persona* concuerda con el número en *Actividad*:

```
select    Edad, Salario
from      Persona
join      Actividad
on        Núm.persona=Núm.actividad
```

Unión exterior (*Outer join*): Selecciona todos los renglones en *Persona* y agrega la información concordante de *Actividad*, si existe:

```
select        Edad, Salario
from          Persona
left outer join  Actividad
on            Persona.núm=Actividad.núm
```

Este sencillo ejemplo nos muestra que el nivel de dificultad técnica o sintáctica, aumenta rápidamente. En ocasiones esto puede volver complicado determinar si el resultado producido por el emparejamiento avanzado es correcto. Un código SQL escrito de manera descuidada o incorrecta puede dar, en el peor de los casos, un resultado aparentemente correcto —la respuesta tiene el número correcto de reglones y los valores en el orden correcto—, pero aun así es incorrecto. En consecuencia, casi siempre será preferible realizar esta tarea siguiendo varios pasos sencillos y controlables, guardando resultados intermedios temporalmente y verificándolos en cada etapa, en vez de hacerlo todo en un paso avanzado gigante. Es más importante tener la certeza de que el resultado será correcto que escribir un código muy elaborado.

Otro método para simplificar un procesamiento complicado consiste en preparar emparejamientos comunes creando *vistas*. Con una *vista*, una expresión compleja de SQL puede ocultarse para que el usuario sólo vea una simple tabla de base de datos que contiene todas las variables de su interés. En nuestro ejemplo, podemos crear de manera definitiva la *vista PersonaActividad*, que contendrá nuestra unión de cuatro tablas. Después sólo nos referiremos a la vista y no a las cuatro tablas cada vez que necesitemos consultar los datos. Por ejemplo, resulta ventajoso convertir las matrices de datos para un "objeto combinado" (sección 9.2) a una *vista*, que podrá servir para recuperar los valores ponderados.

Además de los problemas para escribir la sintaxis que nos permita obtener los resultados que queremos, cuando se hace un emparejamiento a menudo se presenta el problema de la selección del contenido, o sea, elegir los renglones correctos de las tablas de bases de datos. En nuestro ejemplo, podemos suponer que el Registro de Actividades contiene una dimensión temporal. Un emparejamiento sintácticamente correcto arrojaría un resultado sin sentido si, por ejemplo, olvidamos indicar la referencia del punto en el tiempo en que se requiere la información.

Las herramientas adecuadas para el trabajo

En general, en este capítulo hemos partido del supuesto de que los datos están almacenados en una base de datos relacional. Desde el punto de vista técnico, en estos casos se usaría SQL para tener acceso a los datos. Pero esto no se debe interpretar como si todos los usuarios tuvieran que hacer todo el procesamiento con SQL directamente en las bases de datos, sino que a menudo resulta conveniente recurrir a una herramienta que ofrezca una interfaz más fácil de usar (hacer clic, elegir de la lista, arrastrar y soltar, etc.) para recopilar y manipular datos, y que ofrezca un soporte adecuado para el emparejamiento.

Por otro lado, no siempre es verdad que el método más efectivo sea dejar que el sistema de gestión de bases de datos lleve a cabo todo el trabajo. Este sistema está optimizado para trabajar *según la cantidad*. Ése es su método de trabajo normal, y muy eficaz por cierto. Es posible forzar al sistema de gestión de bases de datos a trabajar *secuencialmente* para que realice tareas que requieren ese mecanismo —es decir, seleccionar muestras—, pero con

este método (avanzar con un *cursor*) el sistema se vuelve ineficaz. En esos casos, y en muchos otros en los que se necesita un procesamiento estadístico como la estimación con ponderadores, sería más apropiado usar, por ejemplo, SAS.

Por desgracia, hay muchas situaciones en las que no queda del todo claro cuál es la herramienta más apropiada. Las herramientas de uso más popular tienen una cantidad considerable de áreas de aplicación comunes. Muchas tareas se pueden llevar a cabo de tres o cuatro maneras distintas con otras tantas herramientas. La habilidad y la experiencia del usuario suelen ser más importantes para el resultado que la conveniencia objetiva de la herramienta para realizar una tarea específica de manera rápida y segura. La desventaja es que se corre el riesgo de que las soluciones dependan de la persona.

CAPÍTULO 12

Protección de la privacidad y la confidencialidad

Hasta 1990 los censos de población en Suecia fueron censos "tradicionales" mediante el envío de cuestionarios a cada hogar. Desde el Censo de Población y Vivienda de 1970, con cada nuevo censo de población se han desatado debates sobre la privacidad que también han hecho que aumente la no respuesta en encuestas como la Encuesta sobre Fuerza Laboral. Esto demuestra que la preocupación por la privacidad puede representar una amenaza para las operaciones de una oficina nacional de estadística. Cabe señalar que incluso en Suecia, donde hay una larga tradición de acceso a información pública resguardada por las autoridades, existe esta preocupación.

En la actualidad hay muchas amenazas para la privacidad: cada vez que hacemos una llamada telefónica o usamos nuestra tarjeta de crédito, queda un registro, y en las calles de muchas ciudades hay cámaras que nos siguen cuando caminamos. Aunque estos nuevos tipos de registro no tienen nada que ver con las oficinas de estadística y hoy se divulgan menos los registros estadísticos, es importante que las oficinas de estadística conserven altos estándares de protección de la privacidad tanto de las personas como de las empresas.

Si bien, comparativamente, el riesgo real de que el aprovechamiento estadístico de los datos amenace la privacidad de las personas es reducido, un debate en los medios de comunicación podría traer graves consecuencias para la reputación de una oficina de estadística.

Usamos el término *confidencialidad*, que se aplica tanto a las personas como a las empresas. Cuando se mejora el sistema de registros para facilitar la integración de fuentes y producir estadísticas de mayor calidad, también se debe fortalecer la protección de la confidencialidad. Esto se puede lograr de las siguientes maneras:

1. Minimizando la existencia de variables con *texto*.
2. Minimizando la existencia de variables de identificación como *números de identificación oficial*.
3. Minimizando el riesgo de que puedan derivarse datos sobre personas o empresas de los *tabulados estadísticos* incluidos en publicaciones o bases de datos oficiales.
4. Procesando las matrices que contienen *microdatos* para reducir al mínimo los riesgos de que se pueda conocer la información de las personas, antes de dar el acceso a los investigadores que las analizan.

De acuerdo con los incisos (1) y (2) anteriores, la protección de la confidencialidad puede fortalecerse de dos maneras: reduciendo al mínimo tanto el número de bases de datos con texto o números de identificación oficial, como el número de personas que tienen acceso a esas bases de datos.

La confidencialidad de la información que se suministra a las oficinas de estadística para producir estadísticas está amparada por la legislación de cada país. Además, dentro de una oficina de estadística el procesamiento de los datos se debe organizar de tal modo que se proteja eficientemente la confidencialidad. En este capítulo abordamos la seguridad interna de los datos de los registros, y los riesgos de revelar información confidencial relacionados con la publicación de tabulados y los microdatos que usan los investigadores.

El control estadístico de la revelación de datos confidenciales es un campo vasto y de gran interés que han descrito, por ejemplo, el Federal Committee on Statistical Methodology (2005) y Willenborg y de Waal (2001).

12.1 SEGURIDAD INTERNA

Se debe minimizar el número de bases de datos con nombres, domicilios y otra información en texto. También se puede minimizar el uso de números de identificación oficial —por ejemplo, números de seguridad social o claves sociales o fiscales— si la oficina de estadística recodifica estas claves oficiales como números de identificación para sus registros. Si además se minimiza el número de personas con acceso a estas bases de datos, la seguridad interna estará orientada a proteger la confidencialidad.

12.1.1 No debe haber texto en las bases de datos de salida

Los datos administrativos entregados a la oficina de estadística pueden contener nombres, domicilios y otros datos en forma de texto cuando la autoridad administrativa así lo requiere para sus propósitos. Los registros salariales de los empleadores también contienen datos en lenguaje común. En el procesamiento preliminar de los registros, estos detalles se deben remplazar por códigos.

El siguiente es un ejemplo ficticio:

Gráfico 12.1 Registro administrativo: datos no procesados en la base de datos de entrada

NIP	Nombre	Domicilio	Código postal	Empresa, establecimiento	Puesto	Código TNS	Salario real	Jornada de trabajo
560230123	Pson Per	Calle 1 núm. 7	111 11	Oficina de Estadística de Suecia Estocolmo	Especialista en informática	4321	18 340	0.60
670631234	Ason Eva	Calle 2 núm. 2	777 77	Oficina de Estadística de Suecia Örebro	Jefe de departamento	1234	45 780	1.00

Como parte de la validación de los datos, el número de identificación personal (NIP) se debe suministrar a la oficina de estadística en el formato que ésta indique, todo el lenguaje común se debe remplazar por códigos y todas las variables que carecen de interés estadístico se deben eliminar. Además, se importarán las variables estadísticamente relevantes y se formarán las variables derivadas.

En el ejemplo del gráfico 12.1, el NIP sueco consiste en la fecha de nacimiento (aammdd) más cuatro dígitos individuales. El formato correcto en la Oficina de Estadística de Suecia también incluye el año con cuatro dígitos, como vemos en el gráfico 12.2. Los nombres y los números de identificación de las personas del gráfico 12.1 se cotejan con el Registro de Población y a partir de ese momento ya no es necesario el nombre, pues en la Oficina de Estadística de Suecia todos los emparejamientos se llevan a cabo con la variable NIP.

En el gráfico 12.2, el domicilio de la vivienda y del establecimiento que aparecen en el gráfico 12.1 se remplazan por códigos regionales, y los códigos para puesto y ocupación (TNS) por los códigos ocupacionales CIUO (véase sección 6.2.2). También se crea una variable para salario de tiempo completo y se importa el código de nivel de escolaridad del Registro de Educación.

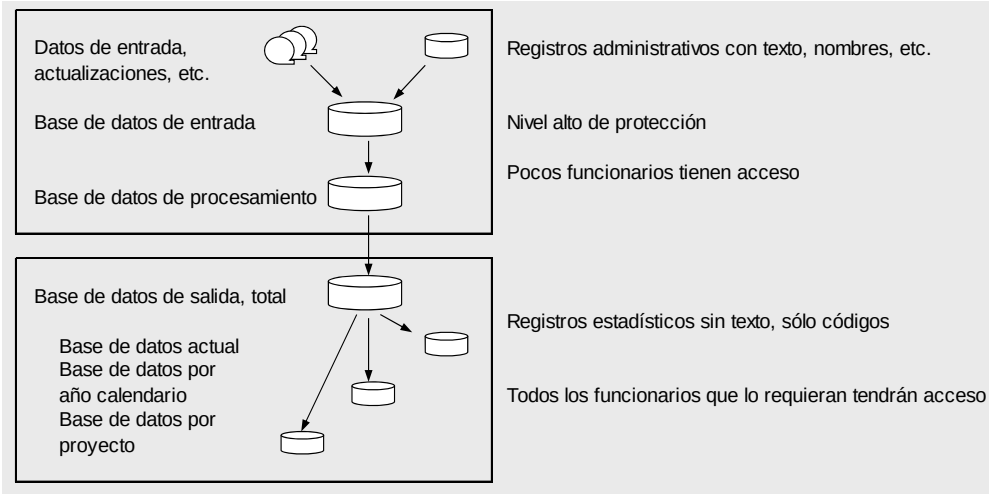
Gráfico 12.2 Registro estadístico correspondiente: datos procesados en la base de datos de salida

NIP	Municipio de residencia	Núm. de identificación de establecimiento	Municipio del establecimiento	Ocupación (CIUO)	Código de nivel de escolaridad	Salario real	Jornada de trabajo	Salario de tiempo completo
19560230123	0180	12345678	0180	2222	1234567	18 342	0.60	30 570
19670631234	1880	23456789	1880	3333	7654321	45 780	1.00	45 780

En este ejemplo, sería recomendable tener acceso al nombre del puesto original en caso de futuros cambios en las clasificaciones ocupacionales, por decir algo. El método más sencillo sería sacar las fuentes administrativas del sistema de datos y archivarlas en un espacio bloqueado una vez que el registro estadístico se haya validado y procesado. Esto permitiría llevar a cabo controles en el futuro sin tener los datos administrativos disponibles en el sistema.

Los nombres, domicilios y otros detalles en lenguaje común, ya sea de personas o empresas, *sólo* deben aparecer en las bases de datos de entrada, por un lado, y en las bases de datos de salida de los registros base, por el otro. Únicamente se permitirá el acceso a ellos a un reducido número de funcionarios de la oficina de estadística, por ejemplo, los que trabajan con entrevistas y cuestionarios. Como regla general, salvo los registros base, las bases de datos de salida no incluirán nombres, texto, etc. y se deberán anonimizar por completo mediante el uso de códigos.

Gráfico 12.3 Diferentes tipos de bases de datos con distintas clases de acceso



12.1.2 Existencia de números de identificación

Además de verificar minuciosamente qué funcionarios estarán autorizados a trabajar con microdatos y minimizar la existencia de variables con texto, el uso de números de identificación de objetos también debe ser restrictivo. Las oficinas de estadística de Dinamarca y

de los Países Bajos han decidido sustituir los números de identificación oficiales de personas por números de identificación de registro, creados y usados por esas dependencias. Esto se hace con la intención de evitar el mal uso de números de seguridad social o de uso fiscal y del NIP, tema tratado en Statistics Netherlands (2004, p. 249).

Número de identificación personal

En primer lugar, debemos reducir el uso de números de identificación personal. Estos números se presentan en las fuentes administrativas que se entregan a la oficina de estadística, ya sea en registros administrativos sobre personas, en algunos casos sobre pequeñas empresas, sobre actividades y en el registro inmobiliario. También debe limitarse el número de funcionarios que reciben estos datos administrativos en la oficina de estadística.

Los datos administrativos se transfieren, luego de la validación y el procesamiento, a las bases de datos de entrada, de procesamiento y de salida. En todas las bases de datos de entrada con números de identificación personal, éstos se deben remplazar por el número de identificación de registro (NIR) de la oficina de estadística, que es único para cada persona. Debe haber una base de datos central, de acceso restringido, donde se realice esta recodificación.

En países donde no hay números de identificación personal, se usa el nombre, el domicilio, la fecha de nacimiento y el lugar de nacimiento para emparejar las entradas de los diferentes registros. Estas variables identificadoras se pueden remplazar por números de identificación de registro de la misma manera en que los NIP oficiales se remplazan con ayuda de una base de datos central.

Después de remplazar estas variables de identificación sensibles, todas las bases de datos de salida acerca de personas se pueden integrar con el NIR. De este modo, la mayoría de los funcionarios que trabajan con estadísticas basadas en registros no tendrá acceso a nombres, etc. o a números de identificación oficial.

Statistics Netherlands (2004, p. 254) menciona otro método que fortalecerá la protección de la confidencialidad. Resulta deseable almacenar datos de registros sobre personas en muchos registros con un número limitado de variables en cada uno. Los funcionarios sólo tienen acceso a registros distintos del que les corresponde cuando necesitan variables de otros registros. De esta forma, se reduce al mínimo el número de funcionarios de la oficina de estadística con acceso a datos confidenciales.

12.2 RIESGOS DE LA REVELACIÓN DE DATOS CONFIDENCIALES: TABULADOS

Si es posible derivar de un tabulado estadístico las características confidenciales de personas o empresas, la publicación de ese tabulado conlleva el riesgo de revelación de datos confidenciales, lo que significa que es posible identificar objetos o valores de variables para objetos individuales.

En el caso de las encuestas por muestreo en las que no se seleccionan los estratos completos, el riesgo de revelación de datos confidenciales suele ser bajo, pero los tabulados basados en censos o en encuestas basadas en registros siempre se deben verificar para que los detalles sobre los objetos individuales no se puedan revelar.

12.2.1 Reglas para tabulados con conteos, totales y valores medios

En el tabulado siguiente, se muestran datos ficticios relacionados con empresas en una rama industrial y un municipio específicos.

Gráfico 12.4 Empresas por tamaño dentro de la rama industrial Y del municipio X

Número de empleados	Número de empresas	Número de empleados	Facturación (miles de coronas)	Sumas salariales (miles de coronas)	Facturación promedio por empresa	Salario anual promedio por empleado
	(1)	(2)	(3)	(4)	(5) = (3) / (1)	(6) = (4) / (2)
0–9	9	50	28 250	11 800	3 139	236
10–99	5	190	116 900	43 380	23 380	228
100–199	3	615	391 650	151 200	130 550	246
200–499	2	600	287 000	169 400	143 500	282
500–	1	705	240 550	159 400	240 550	226
Total	**20**	**2 160**	**1 064 350**	**535 180**	**53 218**	**248**

Se pueden aplicar diferentes reglas para decidir si el riesgo de que algunas celdas de un tabulado revelen datos confidenciales es demasiado elevado. Más adelante ejemplificamos tres reglas con datos del tabulado anterior. Si las celdas tienen un riesgo demasiado elevado de revelación de datos confidenciales de acuerdo con una regla, no se deben publicar.

Tampoco se puede rediseñar el tabulado con menos intervalos de clase, ni suprimir algunas celdas. Existen otras alternativas; el tabulado se puede modificar de diversas maneras descritas en Federal Committee on Statistical Methodology (2005), y Willenborg y de Waal (2001).

En dos renglones del tabulado anterior se revelan datos sobre dos empresas:

– En el renglón con una sola empresa, pues ésta se puede identificar por las variables *municipio*, *rama industrial* y *tamaño*.

– El renglón que contiene dos empresas con 200 a 499 empleados. La gente que conoce una de las dos empresas podría identificar los detalles de la otra mediante una simple resta.

Por consiguiente, se deben suprimir los datos de las columnas (2)–(6) de estos dos renglones. A continuación presentamos tres reglas frecuentes para determinar si una celda es sensible, es decir, si supone un alto riesgo de revelación de datos confidenciales. Estas reglas se pueden combinar en una regla más compleja.

1. Regla del umbral

Una celda de un tabulado se puede definir como sensible si se basa en menos de k observaciones. Pero no sólo se debe suprimir esa celda, sino también otras que permiten calcular las frecuencias de esa celda.

Ejemplo: Con $k = 3$ los dos renglones finales del gráfico 12.4 son sensibles. Si se considera que la columna (1) no revelará información confidencial, podrá publicarse, pero no las otras columnas.

2. Regla (n, k) o de dominancia

Una celda de un tabulado se puede definir como sensible si los valores n más altos suman al menos $100k$% del total de la celda. Esta regla se aplica a las celdas con totales o con sumas.

Ejemplo: De acuerdo con la regla (1, 0.75), la última línea del gráfico 12.4 es sensible porque la empresa representa 100% de todas las sumas. Hay dos empresas que tienen de 200 a 499 empleados. Si la mayor de ellas tiene los atributos mostrados en el gráfico 12.5,

la columna (4) en el renglón de las dos empresas con 200 a 499 empleados es sensible, porque 137,000/169,400 = 0.80, así que no se deben publicar las columnas (4) y (6) de ese renglón.

La columna (3) no es sensible porque 195/287 = 0.68, de modo que las columnas (3) y (5) del gráfico 12.4 sí se pueden publicar de acuerdo con esta regla.

Gráfico 12.5 Un renglón del gráfico 12.4

Número de empleados	Número de empresas	Número de empleados	Facturación (miles de coronas)	Sumas salariales (miles de coronas)	
	(1)	(2)	(3)	(4)	
200–499	2	600	287 000	169 400	← ¿Se pueden publicar estas celdas?
La mayor de estas dos:		400	195 000	137 000	

3. La regla de porcentaje de p

Una celda de un tabulado se puede definir como sensible si es posible estimar el valor de cuando menos un objeto en la celda con un error inferior a p%. Esta regla se aplica a las celdas con totales o sumas. En el ejemplo siguiente, se ilustra la versión más sencilla de esta regla.

Ejemplo: Si los dos últimos renglones del gráfico 12.4 se combinan en un intervalo de clase, con 200 o más empleados, obtenemos los microdatos del gráfico 12.6 para estas tres empresas. Suponemos que una persona que trabaja en la empresa mediana trata de estimar los valores correspondientes a la empresa más grande.

Gráfico 12.6 Los dos últimos renglones del gráfico 12.4 se combinan en un solo renglón

Número de empleados	Número de empresas	Número de empleados	Facturación (miles de coronas)	Sumas salariales (miles de coronas)	
	(1)	(2)	(3)	(4)	
200–	3	1 305	527 550	328 800	← ¿Se pueden publicar estas celdas?
Empresa más pequeña		200	92 000	32 400	
Empresa mediana		400	195 000	137 000	
Empresa más grande		705	240 550	159 400	

Se usará la regla de porcentaje de p con p = 15%. Una persona con información sobre la empresa mediana puede estimar los valores acerca de la empresa más grande de la siguiente manera:

Número de empleados = 1 305 – 400 (conocido) – 200 (límite inferior) = 705
705/705 = 1.00, error inferior a 15%: no se debe publicar.

Facturación aproximada = 527 550 – 195 000 (conocido) = 332 550
332 550/240 550 = 1.38, error mayor a 15%: sí se puede publicar.

Suma salarial aproximada = 328 800 – 137 000 (conocido) = 191 800
191 800/159 400 = 1.20, error mayor a 15%: sí se puede publicar.

La regla de porcentaje de p con p = 15% indica que se pueden publicar las celdas (1), (3) y (4).

12.2.2 Regla del umbral: analice tabulados completos

En el tabulado siguiente se muestra el número de personas convictas en cierta ciudad y cierto año. Algunas celdas con información sensible también contienen números pequeños.

Gráfico 12.7 Ciudad Z, número de personas convictas por edad y sexo

Edad	Mujeres	Hombres	**Ambos sexos**
16–24	0	7	**7**
25–34	1	7	**8**
35–44	3	17	**20**
45–54	3	3	**6**
55–64	3	5	**8**
Total	**10**	**39**	**49**

Con los datos de este tabulado, no es posible derivar información confidencial de ninguna persona. Pero sí es posible hacerlo si se tiene la información del número de personas de la ciudad Z, por edad y sexo.

De acuerdo con la regla del umbral con $k = 3$, se puede suponer que las celdas del tabulado marcadas con color gris son publicables. Pero se trata de una interpretación incorrecta de la regla, pues deben ser las celdas del tabulado *completo* las que deben corresponder al menos a tres observaciones. La base de un tabulado completo es el número total de personas de la población.

El gráfico 12.7 es en realidad sólo una parte de un tabulado de frecuencias más grande y completo en el que la población se divide entre tres variables: sexo, edad y delito, medidos por las categorías convicto y nunca convicto.

En el tabulado completo, la población se tabula simultáneamente por sexo, edad y delito. Ahí es donde se pueden observar los riesgos de revelación de datos confidenciales.

Gráfico 12.8 Ciudad Z, población por sexo, edad y delito

Sexo	**Edad**	**Personas convictas**	**Personas no convictas**	**Población total**	**Riesgo de revelación de datos**
Mujeres	16–24	0	2	2	0.00
	25–34	1	1	2	0.50
	35–44	3	28	31	0.10
	45–54	3	41	44	0.07
	55–64	3	5	8	0.38
Total de mujeres	*16–64*	*10*	*77*	*87*	
Hombres	16–24	7	0	7	1.00
	25–34	7	1	8	0.88
	35–44	17	65	82	0.21
	45–54	3	65	68	0.04
	55–64	5	3	8	0.63
Total de hombres	*16–64*	*39*	*134*	*173*	
Ambos sexos	**16–64**	**49**	**211**	**260**	

En el tabulado completo, queda claro qué partes no son adecuadas para publicación. La regla del umbral del tabulado incompleto del gráfico 12.7 con al menos tres como base significa que las celdas blancas no se deben publicar. Cuando vemos a una mujer entre 16 y 24 años, se puede llegar a la conclusión de que no ha sido convicta. Tal vez esa información no sea tan delicada, pero a una mujer entre 25 y 34 años la regla del umbral realmente la protegería contra la revelación de información confidencial.

Sin embargo, en el tabulado completo del gráfico 12.8 hay dos celdas adicionales con frecuencias menores a 3. Las celdas con 0 y 1 hombres no convictos se deben suprimir, pero también las celdas con 7 y 7 hombres convictos, porque las celdas suprimidas se pueden calcular a partir de éstas y el valor de la población total.

Esto significa que no se deben publicar las celdas para hombres de 16 a 24 años y de 25 a 34 años. Se deben suprimir incluso más celdas, pues es posible hacer buenas estimaciones de las celdas con información sensible a partir de las otras celdas y los márgenes conocidos. Con todo, no bastan las reglas, sino que también es necesario aplicar nuestro criterio. La última columna del gráfico 12.8 muestra la probabilidad de que una persona descrita por categoría de edad y sexo sea convicta. Estos riesgos pueden usarse para detectar celdas con datos sensibles donde es necesario aplicar nuestro criterio.

12.2.3 La combinación de tabulados puede revelar datos confidenciales

En el sitio web de una oficina de estadística se pueden consultar tabulados detallados que se basan en registros. Si, por ejemplo, se combina un tabulado para una región pequeña con un tabulado obtenido para un proyecto especial que incluye datos para mapas de polígonos del sistema de información geográfica (SIG) de la oficina de estadística, puede ser posible una revelación de datos confidenciales. Si esa región pequeña consiste en un polígono del SIG más un bosque más una pequeña población, se obtienen los siguientes tabulados:

Gráfico 12.9 Tabulado para la región

Ocupación	Sexo	Número de personas	Ingreso promedio
Estadístico			
	F	8	32 560
	M	7	32 780
...	...	...	...

Gráfico 12.10 Tabulado para el polígono del SIG

Ocupación	Sexo	Número de personas	Ingreso promedio
Estadístico			
	F	7	29 870
	M	6	30 120
...	...	...	...

Combinando estas tabulados es fácil calcular el ingreso de los dos estadísticos que viven en una pequeña población no incluida en el polígono del SIG. Los funcionarios que entregan los datos SIG a los clientes deben considerar que los tabulados se pueden combinar de esta manera. Esto significa que los datos SIG se deben verificar cuidadosamente antes de su entrega.

Los tabulados producidos por una oficina de estadística también se pueden combinar con tabulados elaborados, por ejemplo, por una asociación gremial. La oficina de estadística publica datos para toda la rama industrial en cuestión, pero la asociación produce tabulados sólo para sus miembros. Si todas las empresas grandes menos una pertenecen a la asociación, será fácil calcular los valores para la empresa que no es miembro. Este tipo de posibilidad es muy difícil de prevenir.

12.3 RIESGOS DE LA REVELACIÓN DE DATOS CONFIDENCIALES: MICRODATOS

Muchos registros estadísticos son de gran importancia para investigaciones en las áreas de medicina o ciencias sociales. Un gran número de investigadores usa registros que combinan datos de diferentes campos, así como registros longitudinales. A las oficinas de estadística les toca la importante tarea de hacer que esos datos estén disponibles para propósitos de investigación.

Los registros usados por investigadores deben anonimizarse. Cuando se entrega un registro a investigadores, se debe examinar qué variables deben borrarse. En los registros sobre personas, se deben eliminar nombre, domicilio, número de identificación personal y código inmobiliario. En los registros sobre empresas, el sector industrial, el tamaño y la región

también pueden ser variables sensibles que podrían permitir la identificación de las empresas más grandes.

Hay varias maneras de minimizar el riesgo de revelar datos confidenciales, por ejemplo:

– Limitar las variables que incluidas en los datos que se entregan a los investigadores.
– Agrupar las variables de expansión en la menor cantidad posible de categorías, es decir, en vez de dividir los datos por parroquias, sólo se dividirán por condados.
– Permitir a los investigadores el acceso a una muestra de la población de registro actual. Así se debe iniciar cualquier colaboración. Si el investigador tiene la oportunidad de conocer primero el material, entonces podrá crear una matriz de datos más grande que sólo incluya las variables de su interés.

No obstante, la mejor manera de proteger la confidencialidad de los microdatos consiste en usar una combinación de acceso remoto, legislación y acuerdos o convenios para otorgar las licencias para el uso de la información.

El acceso remoto significa que los investigadores sólo pueden analizar microdatos anonimizados en su lugar de trabajo vía internet. En este caso, los microdatos siempre permanecen en la oficina de estadística. Los investigadores no tienen acceso a los microdatos, pero pueden aplicar técnicas estadísticas a los datos y obtener sólo los resultados del análisis ejecutado.

Quienes obtienen permiso para analizar microdatos deben asumir cuando menos las mismas obligaciones legales que los funcionarios de la oficina de estadística para proteger a las personas y las empresas, y garantizar que no se revele información confidencial. Los investigadores sólo obtendrán ese permiso después de celebrar un acuerdo o convenio jurídicamente vinculante.

CAPÍTULO 13

Coordinación y coherencia

Para las encuestas basadas en registros, es importante que la oficina de estadística cuente con una organización que apoye una amplia colaboración entre los responsables de las distintas partes del sistema de registros. Los administradores desempeñan un importante papel en esta coordinación, entre otras cosas, para asegurar el funcionamiento conjunto de las diferentes partes del sistema de registros. Sin embargo, no basta con una organización formal apropiada. Debe haber una cultura organizacional donde sea inherente la colaboración más allá de las fronteras de la organización.

Quienes trabajan con estadísticas basadas en registros deben adoptar un enfoque común y no pensar sólo en su registro. También deben entender su función dentro del sistema. Por un lado, está la responsabilidad de apoyar el trabajo de otros registros y, por el otro, la posibilidad de usar el sistema para desarrollar nuevas aplicaciones.

Para lograr esa comprensión, es importante ofrecer capacitación, seminarios, proyectos conjuntos y tareas. También debe alentarse la movilidad dentro de la organización. Constituye una ventaja tener un gran número de funcionarios con experiencia en varios registros dentro del sistema.

La colaboración antes mencionada es una condición fundamental cuando se llevan a cabo registros y encuestas coordinados. El propósito de la coordinación es obtener estimaciones coherentes, para lo que se requieren registros coordinados y métodos de estimación adecuados. En este capítulo abordaremos esa coordinación y los métodos de estimación aplicados en todo el sistema.

13.1 COORDINACIÓN RELACIONADA CON LOS CONTENIDOS

Para asegurar que los datos estadísticos de diferentes encuestas sean comparables, es necesario *definir* de manera consistente las poblaciones y variables de esas encuestas. Además, los registros se deben *procesar* con el propósito de lograr consistencia entre las poblaciones de registro y las variables de los diferentes registros estadísticos. La coordinación del contenido se basa en los siguientes elementos:

– *poblaciones* estandarizadas
– *variables* estandarizadas

13.1.1 Coordinación y estandarización de las poblaciones

En el capítulo 5 se describe la creación de poblaciones de registro y la definición de objetos. Todas las poblaciones de registro deben crearse usando los registros base. Una aplicación importante de los registros base es la creación de *poblaciones estandarizadas*, que después se utilizan para crear las poblaciones de registro en otros registros del sistema.

Estas poblaciones estandarizadas deben ser de alta calidad en lo que respecta a la cobertura y las variables de expansión relevantes. Por lo tanto, se crean *después* del periodo o momento al que se refieren. En la medida de lo posible, se debe reportar y procesar toda la información administrativa antes de definir la población. No debe modificarse el registro que contiene las poblaciones estandarizadas. De este modo, si otros registros usan exclusivamente esas poblaciones de registro estandarizadas, las estadísticas basadas en registros serán consistentes en cuanto a los conjuntos de objetos.

Varios registros suecos tienen como punto de partida la población estandarizada del Registro de Población que se refiere al 31/12 de cada año. En la sección 5.6 explicamos cómo producir estadísticas regionales basadas en registros detallados y completamente consistentes.

No obstante, la consistencia es más problemática en el caso de las estadísticas económicas. En Suecia, las encuestas de empresas se basan, en buena medida, en una captación de datos propia por medio de poblaciones marco formadas en el mes de noviembre previo al periodo de la encuesta.

En las secciones 5.4.8 y 5.4.9 describimos en términos generales los problemas que surgen con las poblaciones marco. En el ejemplo siguiente vemos la manera en que las inconsistencias en la población afectan las estadísticas sobre empresas de energía.

Gráfico 13.1 Poblaciones marco inconsistentes

Se preparan encuestas sobre el ingreso y el gasto de las empresas de energía suecas, tanto dentro de las Estadísticas de Empresas (EEm) como en las Estadísticas de Energía (EEn). La validación de la consistencia ha demostrado que se incluyeron diferentes números de identificación en las dos poblaciones marco. Algunos números de identificación se incluyen en ambas encuestas y otros en una sola.

Ingreso en 1998 según:		
	EEm	EEn
Sólo EEm	32%	0%
Ambas	68%	76%
Sólo EEn	0%	24%
Total	100%	100%

Ingreso en 1999 según:		
	EEm	EEn
Sólo EEm	16%	0%
Ambas	84%	96%
Sólo EEn	0%	4%
Total	100%	100%

Costos en 1998 según:		
	EEm	EEn
Sólo EEm	32%	0%
Ambas	68%	84%
Sólo EEn	0%	16%
Total	100%	100%

Costos en 1999 según:		
	EEm	EEn
Sólo EEm	18%	0%
Ambas	82%	96%
Sólo EEn	0%	4%
Total	100%	100%

Como los problemas con las poblaciones marco son significativos, se considera deseable usar para muchas encuestas una versión más actual del Registro Empresarial. Dado que diferentes encuestas se basan en ese registro en distintos momentos, las poblaciones de estas encuestas también serán diferentes. También puede variar la clasificación industrial para la misma empresa, lo que significa que la población de empresas de energía podría diferir dependiendo de la encuesta.

13.1.2 Coordinación y estandarización de variables

En las estadísticas económicas, algunas inconsistencias se deben a diferencias en la definición de las variables y los métodos de medición. El ejemplo del gráfico13.2 nos presenta una descripción de la manera en que las empresas de energía suecas se ven afectadas por inconsistencias relacionadas con las variables. Los valores medidos para el ingreso y el gasto de cada empresa de energía pueden diferir entre las dos encuestas.

Gráfico 13.2 Variables inconsistentes

Rama industrial (véase sección 9.2.2)

Las Estadísticas de Empresas (EEm) presentan sus informes de acuerdo con la rama industrial *principal*, es decir, truncan la variable de valores múltiples de rama industrial, de modo que el ingreso y el gasto por concepto de energía pueden presentarse en la rama industrial "equivocada".

Las Estadísticas de Energía (EEn) presentan sus informes de tal modo que las empresas activas en varias ramas industriales sólo informan del gasto e ingreso por concepto de energía a la rama industrial energética.

Ingreso y gasto

Empresas incluidas con el mismo número de identificación en las estadísticas de Empresas y de Energía:

Ingreso 1999	**Costos 1999**
según la EEm: 108 mil millones	según la EEm: 64 mil millones
según la EEn: 103 mil millones	según la EEn: 51 mil millones
Empresa con la mayor diferencia:	Empresa con la mayor diferencia:
EEn – EEm = 2 000 millones	EEm – EEn = 6 000 millones

13.2 COHERENCIA

El concepto de *consistencia* tiene que ver con el acuerdo. Dos encuestas son consistentes si se relacionan con la misma población y si los objetos tienen los mismos valores de medición para las variables comunes, con las mismas estimaciones.

El concepto de *coherencia* se refiere al hecho de que las estimaciones de diferentes encuestas se pueden usar conjuntamente. Por ejemplo, para que un cociente tenga sentido, el numerador y el denominador deben ser coherentes. Las encuestas consistentes arrojan estimaciones coherentes.

La coherencia, que se puede mejorar mediante coordinación y métodos de estimación adecuados, se considera un componente de la calidad de la información estadística en muchas oficinas de estadística.

La Oficina de Estadística de los Países Bajos adoptó un nuevo enfoque en el que se da prioridad a la consistencia de las estadísticas (Statistics Netherlands, 2000). Las estadísticas de fuentes diversas pueden ser consistentes, es decir, tener un alto nivel de coherencia, de acuerdo con la metodología de los Países Bajos:

– asegurando la consistencia de las *poblaciones* (en lo concerniente a las definiciones tanto de objetos como del conjunto de objetos);
– asegurando la consistencia respecto a las *variables*;
– usando métodos de calibración que produzcan *estimaciones* consistentes.

Para asegurar que las estadísticas sean consistentes, no nos debemos limitar a las estadísticas basadas en registros, sino también incluir todo tipo de encuestas: se deben coordinar las encuestas por muestreo, los censos y las encuestas basadas en registros.

Consistencia de las poblaciones

Para lograr consistencia en las poblaciones, se deben coordinar los conjuntos de objetos para todas las encuestas relativas a un tipo de objeto específico y una referencia temporal. Esto se puede hacer creando una población de registro común para todas las encuestas con el mayor nivel de calidad posible, de modo que sólo se debe crear una vez recibida toda la información relevante de la oficina de estadística.

Consistencia de las variables

Se deben coordinar las definiciones de las variables. Las variables relacionadas se deben validar en conjunto y se deben hacer las mismas correcciones e imputaciones en todas las encuestas.

Métodos de calibración para una estimación consistente

El método, descrito por Houbiers *et al*. (2003), consiste en hacer calibraciones repetidas que dan estimaciones consistentes para todas las encuestas de que se trate. Los objetos de la población común se dividen en una serie de bloques donde todos los objetos de cada bloque tienen valores para las mismas variables. Después se forman estimaciones para cada bloque.

Se empieza por un bloque con información de registro para todos los objetos, luego se hacen estimaciones para el siguiente mejor bloque —en lo que respecta a calidad— consistentes con las estimaciones del registro. Las estimaciones para el segundo bloque se calculan mediante el método de calibración descrito en la sección 7.5. El siguiente paso es hacer estimaciones para el tercer bloque que sean consistentes con las del bloque precedente, y así sucesivamente.

13.3 ESTADÍSTICAS DE EMPRESAS CONSISTENTES Y COHERENTES

Usaremos el ejemplo de una empresa a fin de ilustrar un método para producir estadísticas coordinadas y consistentes. La finalidad es mostrar las inconsistencias que pueden ocurrir en encuestas realizadas en diferentes momentos. Si los datos administrativos que describen cambios en la población se presentan con retraso, las encuestas realizadas tempranamente presentarán errores de marco considerables. Con el método que presentamos aquí, es posible calcular estimaciones corregidas para todas las encuestas basadas en un registro por año calendario.

Todas las encuestas del ejemplo se refieren al año *t*, en el que se llevan a cabo diversas encuestas en diferentes momentos. En primer lugar, se realizan encuestas trimestrales, a partir de la población marco establecida en noviembre del año *t-1*. Esta encuesta se refiere a la producción en la industria manufacturera expresada como valor agregado. Las estimaciones se calculan usando la variable *número de empleados*, una variable auxiliar presente en el marco de muestreo.

A continuación, al inicio del año *t+1*, se lleva a cabo un censo con captación de datos propia, basado en la población marco definida en noviembre del año *t*. Se estudia el valor agregado de todas las empresas. Por último, se realiza una encuesta basada en registros utilizando el Registro de Estados de Resultados. Estos datos, correspondientes al año *t*, estarán disponibles en la primavera del año *t+1*. En el gráfico siguiente se observa el momento en el que se definen los marcos y las poblaciones de registro y se llevan a cabo las diferentes encuestas. En todas las encuestas el tipo de objeto es unidad jurídica (UJ).

Gráfico 13.3 Orden cronológico de los marcos y las encuestas

Noviembre año *t-1*	Se define una población marco (llamada marco I) con la versión de inventario actual del Registro Empresarial, que contiene todas las empresas (unidades jurídicas) conocidas en ese momento.
Abril año *t*	Encuesta por muestreo usando el marco I de empresas pertenecientes a la industria manufacturera (D) sobre el valor agregado durante el primer trimestre del año *t*.
Julio año *t*	Encuesta por muestreo usando el marco I de empresas pertenecientes a la rama industrial D sobre el valor agregado durante el segundo trimestre del año *t*.

Octubre año *t*	Encuesta por muestreo usando el marco I de empresas pertenecientes a la rama industrial D sobre el valor agregado durante el tercer trimestre del año *t*.
Noviembre año *t*	Se define una población marco (llamada marco II) con la versión del inventario actual del Registro Empresarial, que contiene todas las empresas (unidades jurídicas) conocidas en ese momento.
Enero año *t+1*	Encuesta por muestreo usando el marco I de empresas pertenecientes a la rama industrial D sobre el valor agregado durante el cuarto trimestre del año *t*.
Enero año *t+1*	Censo usando el marco II de empresas en todas las ramas industriales sobre valor agregado en el año *t*.
Primavera año *t+1*	Encuesta basada en registros para el año *t* usando los estados de resultados.
Otoño año *t+1*	Se define una población por año calendario para el año *t* con todas las unidades jurídicas activas durante el año *t* a partir de toda la información administrativa disponible.
Otoño año *t+1*	Se corrigen todas las encuestas en lo relativo a la información presente en el registro por año calendario.

13.3.1 Poblaciones inconsistentes

El Registro Empresarial recibe con bastante retraso la información sobre reestructuración, apertura y cierre de empresas. Además, los datos sobre la clasificación industrial de las empresas también pueden ser incorrectos en el marco de noviembre, lo que se detectará posteriormente.

En algún momento del otoño siguiente al año en cuestión, se recibe casi toda la información sobre el año anterior y se puede crear un registro por año calendario de buena calidad. En el gráfico siguiente, se ilustran los errores de marco de los marcos I y II mediante una comparación con la población final de año calendario para el año *t*, que se basa en toda la información disponible durante el otoño del año *t+1*.

Gráfico 13.4 Marcos de muestreo I y II, versión de año calendario del Registro Empresarial

Marco I:
Noviembre del año *t-1*

Id. Emp.	Rama industrial	Núm. de empleados
UJ1	A	5
UJ2	D	210
UJ3	D	40
UJ4	D	120
UJ5	D	15
UJ6	D	30
UJ7	E	55
UJ8	E	70
UJ9	F	90
UJ10	F	5
UJ11	G	340
UJ12	G	15
UJ13	G	10
UJ14	H	60
UJ15	K	20
UJ16	K	5

Marco II:
Noviembre del año *t*

Id. Emp.	Rama industrial	Núm. de empleados
UJ1	A	5
UJ2	D	205
UJ3	D	45
UJ4	D	110
UJ6	D	25
UJ7	E	60
UJ8	E	65
UJ9	F	80
UJ10	F	10
UJ11	G	330
UJ12	G	20
UJ13	G	5
UJ14	H	70
UJ15	K	15
UJ16	K	10
UJ17	**D**	**20**

Versión de año calendario del Registro Empresarial
Año calendario *t*, creado en el otoño del año *t+1*

Id. Emp.	Rama industrial	Núm. de empleados	ponderador
UJ1	A	5	1
UJ2	D	205	1
UJ3	D	45	1
UJ4	D	110	1
UJ5	D	15	0.25
UJ6	D	25	1
UJ7	E	60	1
UJ8	E	65	1
UJ9	F	80	1
UJ10	F	10	1
UJ11	G	330	1
UJ12	G	20	1
UJ13	G	5	1
UJ14	**K**	70	1
UJ15	K	15	1
UJ16	K	10	1
UJ17	D	20	0.5
UJ18	**G**	**10**	**0.5**

El marco I consiste en el inventario actual de las unidades jurídicas activas durante noviembre del año *t-1* de acuerdo con la información disponible en ese momento.

El marco II consiste en el inventario actual de las unidades jurídicas activas durante noviembre del año *t* de acuerdo con la información disponible en ese momento. La unidad jurídica 5 no está activa durante noviembre y se ha agregado la nueva unidad jurídica 17.

El registro por año calendario contiene todas las unidades jurídicas activas durante alguna parte del año *t*. Se crea cuando toda la información administrativa sobre el año *t* está disponible durante el otoño del *año t+1*. La unidad jurídica 5 ha estado activa durante el primer trimestre del año *t*, mientras que las unidades jurídicas 17 y 18 han estado activas durante los últimos dos trimestres del año *t*. Se descubre que la unidad jurídica 14 no pertenece a la rama industrial H, sino a la K. El tiempo que una unidad jurídica se mantiene activa durante el año *t* genera los ponderadores en el registro por año calendario, como se ve en el gráfico 9.2.

Los errores de marco de los marcos I y II son grandes, lo que coincide con nuestra experiencia en el Registro Empresarial sueco (véase gráfico 1.8). Con cinco fuentes creamos una versión de año calendario para el Registro Empresarial que contiene todas las empresas (unidades jurídicas) activas durante 2002. Cada fuente consiste en las unidades jurídicas en un régimen fiscal. En la tabla siguiente, se comparan la subcobertura y la sobrecobertura de las fuentes con nuestra versión final de año calendario del Registro Empresarial.

Solo la fuente 1 estuvo disponible en noviembre de 2001 y se usó como marco para las encuestas por muestreo durante 2002. Los conjuntos de objetos administrativos de cada fuente son adecuados para cada uno de los cinco regímenes fiscales. No obstante, de manera individual, cada fuente es de *baja calidad estadística*, aunque la cobertura será buena si se combinan todas las fuentes en un registro por año calendario.

Gráfico 13.5 Sobrecobertura y subcobertura en cinco fuentes administrativas

	Fuente 1	Fuente 2	Fuente 3	Fuente 4	Fuente 5
Sobrecobertura	41%	0%	0%	0%	0%
Subcobertura	21%	74%	74%	30%	9%

13.3.2 Estimaciones preliminares

El procedimiento es el siguiente:

– Con las encuestas por muestreo (gráfico 13.6), se hacen estimaciones preliminares relacionadas con el valor agregado de la *rama industrial D* durante el año *t*. Con iguales probabilidades, se eligen tres empresas de un total de cinco. Los ponderadores se calibran usando el *número de empleados* como variable auxiliar. Los ponderadores w_i se obtienen calibrando d_i con el *número de empleados*, según se describe en la sección 7.5. La empresa *UJ5* se trata como no respuesta.

– Con el censo (gráfico 13.7), se hacen estimaciones preliminares acerca del valor agregado durante el año *t*.

Gráfico 13.6 Marco de muestreo I y encuesta por muestreo acerca de la rama industrial D, año *t*

A. Marco I: Noviembre de año *t*

Id. de empresa	Rama industrial	Núm. de empleados
UJ1	A	5
UJ2	**D**	**210**
UJ3	**D**	**40**
UJ4	**D**	**120**
UJ5	**D**	**15**
UJ6	**D**	**30**
UJ7	E	55
UJ8	E	70
UJ9	F	90
UJ10	F	5
UJ11	G	340
UJ12	G	15
UJ13	G	10
UJ14	H	60
UJ15	K	20
UJ16	K	5

B. Encuesta por muestreo trimestral, valor agregado en la rama industrial D, año *t*

	Valor agregado (millones de coronas)			
	T 1	T 2	T 3	T 4
UJ2	25	25	22	25
UJ4	15	12	7	14
UJ5	2	no respuesta	no respuesta	no respuesta

C. Ponderadores para diferentes estimaciones

	d_i	T 1: w_i	T 2–4: w_i
UJ2	5/3	1.09	1.18
UJ4	5/3	1.34	1.39
UJ5	5/3	1.63	0.00

D. Valor agregado estimado

T 1	50.7
T 2	46.2
T 3	35.7
T 4	49.0
Total año 2	**181.7**

Dos empresas tienen como resultado no respuesta en el censo del gráfico 13.7. Se les imputaron valores usando el cociente *valor agregado/empleado,* que se calcula a partir de las empresas de la misma rama industrial que han respondido. El valor imputado para *UJ10* se calcula como 10 • 26 / 80 = 3.25 millones de coronas.

Gráfico 13.7 Población marco II y censo, año *t*

E. Marco II: Noviembre de año *t*

Id. de empresa	Rama industrial	Núm. de empleados
UJ1	A	5
UJ2	D	205
UJ3	D	45
UJ4	D	110
UJ6	D	25
UJ7	E	60
UJ8	E	65
UJ9	F	80
UJ10	F	10
UJ11	G	330
UJ12	G	20
UJ13	G	5
UJ14	H	70
UJ15	K	15
UJ16	K	10
UJ17	D	20

F. Censo, año *t*

Valor agregado (millones de coronas)	
Antes imputación	Después imputación
2	2
97	97
21	21
52	52
12	12
93	93
100	100
26	26
no respuesta	**3.25**
112	112
7	7
2	2
60	60
13	13
no respuesta	**8.67**
10	10

G. Estimaciones para el año *t*

Rama industrial	Núm. Emp.	Valor agregado
A	5	2
D	405	**192**
E	125	193
F	90	29.25
G	355	121
H	70	60
K	25	21.67
Total	1 075	618.92

Las estimaciones de la tabla G se hicieron sumando los valores luego de la imputación para cada rama industrial en las matrices de datos E y F. Las estimaciones para la rama industrial D son inconsistentes, pues la encuesta por muestreo da un total anual de 181.7 millones de coronas, mientras que en el censo se estiman 192 millones de coronas.

13.3.3 Estimaciones corregidas y consistentes

En el otoño del año *t+1*, se crea una población estandarizada de empresas que incluye a todas las empresas que han estado activas en cualquier momento del año *t*. Para las empresas de esta población por año calendario, se compilan microdatos de las cuatro encuestas en la misma matriz de datos (gráficos 13.8 y 13.9). Por último, se hacen estimaciones consistentes para el año *t* (gráfico 13.10).

Por lo tanto, la matriz de datos del gráfico 13.8 es completamente *consistente en lo que respecta a la población*. La población de empresas ha cambiado: para empezar, se agregó la empresa *UJ18*, que abrió durante el año, y el código de clasificación industrial de la empresa *UJ14* cambió de H a K. De modo que la matriz de datos del gráfico 13.8 es *consistente en cuanto a la variable de expansión rama industrial*. Los códigos industriales preliminares e incorrectos de los gráficos 13.6 y 13.7 se corrigieron para todas las encuestas.

Además, el registro por año calendario contiene la variable *ponderador*, que indica cuánto tiempo se ha mantenido activa la empresa durante el año. La matriz de datos también contiene sumas salariales basadas en los datos administrativos del Registro de Estados de Resultados. Esta encuesta basada en registros no necesita ninguna corrección en este ejemplo, por lo que no se menciona en la sección 13.3.2.

Persisten ciertos problemas en la matriz de datos del gráfico 13.8:

– Los valores medidos en relación con el valor agregado para la empresa *UJ4* son distintos en la encuesta por muestreo y el censo.

– Hay dos valores imputados en el censo para *UJ10* y *UJ16*; el modelo de imputación se puede mejorar dado que la fuente administrativa con sumas salariales ahora está disponible.

– A causa de los errores del marco, hay dos valores perdidos en el censo (*UJ5* y *UJ18*), que se basa en el marco de noviembre del año *t*.

Gráfico 13.8 Población por año calendario y matriz de datos con todas las encuestas para el año *t*

Población por año calendario del Registro Empresarial				Muestra, valor agregado en año *t* (millones de coronas)					Censo año *t*	Datos administrativos
Id. empresa	Rama industrial	Núm. emplea-dos	ponderador momento del año	T1	T2	T3	T4	Año *t*	Valor agregado (millones)	Suma salarial (millones)
UJ1	A	5	1						2	0.3
UJ2	D	205	1	25	25	22	25	97	97	41.8
UJ3	D	45	1						21	9.2
UJ4	D	110	1	15	12	7	14	**48**	**52**	22.4
UJ5	D	15	0.25	2	perdido	perdido	perdido	perdido	perdido	0.9
UJ6	D	25	1						12	5.1
UJ7	E	60	1						93	13.1
UJ8	E	65	1						100	14.2
UJ9	F	80	1						26	14.5
UJ10	F	10	1						(3.25)	1.8
UJ11	G	330	1						112	51.8
UJ12	G	20	1						7	3.1
UJ13	G	5	1						2	0.8
UJ14	K	70	1						60	13.5
UJ15	K	15	1						13	2.9
UJ16	K	10	1						(8.61)	1.9
UJ17	D	20	0.5						10	2.0
UJ18	G	10	0.5						perdido	0.9

Estos problemas se resuelven de las siguientes maneras:

UJ4:

Hay una inconsistencia en el gráfico 13.8 en la variable valor *agregado*, porque la encuesta por muestreo dio como resultado 48 millones de coronas como valor anual para la *UJ4*, mientras que el censo indica 52 millones de coronas. Suponemos que el censo es más confiable y, en consecuencia, modificamos los valores trimestrales de esta unidad jurídica. Cada valor trimestral se multiplica por 52/48.

UJ5:

Se asigna el valor 0 para los trimestres 2–4 y el valor 2 para el año completo.

UJ10, UJ16 y UJ18:

Los valores perdidos en el censo se remplazan por valores imputados con base en el cociente *valor agregado/suma salarial*, que se calcula usando las empresas de la misma rama industrial que han respondido.

El valor imputado para *UJ10* se calcula como 1.8 • 26 / 14.5 = 3.23 millones de coronas.

El valor imputado para *UJ16* se calcula como 1.9 • (60+13) / (13.5+2.9) = 8.45 millones de coronas.

El valor imputado para *UJ18* se calcula como 0.9 • (112+7+2) / (51.8+3.1+0.8) = 1.96 millones de coronas.

Una vez efectuados estos cambios, la matriz de datos está completa. La matriz de datos final se muestra en el gráfico 13.9.

Gráfico 13.9 Matriz de datos completa con todas las encuestas para el año *t*

Población por año calendario del Registro Empresarial				**Muestra**, valor agregado en año *t* (millones de coronas)					**Censo** año *t*	**Datos administrativos**
Id. empresa	Rama industrial	Núm. empleados	ponderador momento del año	T1	T2	T3	T4	Año *t*	Valor agregado (millones de coronas)	Suma salarial (millones de coronas)
UJ1	A	5	1						2	0.3
UJ2	D	205	1	25	25	22	25	97	97	41.8
UJ3	D	45	1						21	9.2
UJ4	D	110	1	**16.25**	**13.00**	**7.58**	**15.17**	**52**	52	22.4
UJ5	D	15	0.25	2	**0**	**0**	**0**	**2**	**2**	0.9
UJ6	D	25	1						12	5.1
UJ7	E	60	1						93	13.1
UJ8	E	65	1						100	14.2
UJ9	F	80	1						26	14.5
UJ10	F	10	1						**3.23**	1.8
UJ11	G	330	1						112	51.8
UJ12	G	20	1						7	3.1
UJ13	G	5	1						2	0.8
UJ14	K	70	1						60	13.5
UJ15	K	15	1						13	2.9
UJ16	K	10	1						**8.45**	1.9
UJ17	D	20	0.5	Subcobertura en la encuesta por muestreo					10	2.0
UJ18	G	10	0.5						**1.96**	0.9

La matriz de datos del gráfico 13.9 es consistente en lo que respecta a *población* y *variables.*

El siguiente paso es hacer *estimaciones* consistentes para todas las encuestas. El número de empleados de todo el año se estima usando ponderadores que indican cuánto tiempo se mantuvo activa la empresa durante el año (véase sección 9.1). Las sumas salariales y el valor agregado se estiman formando totales para cada industria. Las encuestas por muestreo trimestrales se recalibran (no mostramos los cálculos) de modo que las estimaciones concuerden con la encuesta anual.

Por último, se calculan los cocientes, es decir, la productividad como valor agregado por empleado durante todo el año. Para este cociente se usan numeradores y denominadores de diferentes encuestas. Luego entonces, es una gran ventaja que en la matriz de datos anterior las encuestas sean consistentes en el nivel micro.

Gráfico 13.10 Estimaciones consistentes basadas en una matriz de datos consistente

Rama industrial	Núm. de empleados durante todo el año	Sumas salariales (millones de coronas)	Valor agregado (millones de coronas)	Valor agregado por empleado (millones de coronas)	Salario anual por empleado (miles de coronas)
A	5.0	0.3	2.0	0.40	60.0
D	398.8	81.4	194.0	0.49	204.1
E	125.0	27.3	193.0	1.54	218.4
F	90.0	16.3	29.2	0.32	181.1
G	360.0	56.6	123.0	0.34	157.2
K	95.0	18.3	81.5	0.86	192.6
Total	**1 073.8**	**200.2**	**622.6**	**0.58**	**186.4**

Este ejemplo muestra los principios para producir estadísticas consistentes sobre empresas. Sin embargo, es apenas el primer paso para lograr estadísticas consistentes y coherentes. El número de empleados durante todo el año en el Registro Empresarial también debe ser consistente con el Registro de Empleo, mientras que las sumas salariales deben serlo respecto a los registros de Estados de Resultados y de Ingresos.

CAPÍTULO 14

Conclusiones

Los capítulos anteriores de este libro contienen muchas propuestas de cambio. Se introducen nuevos términos y métodos con la intención de que los sistemas de registros y las estadísticas basadas en registros se desarrollen y funcionen mejor que hoy. Dillman (1996) se muestra más bien pesimista respecto a la innovación y el cambio en las dependencias gubernamentales dedicadas a las encuestas, en especial en el caso de las encuestas que no son por muestreo. Entre otras razones, esto se debe al distanciamiento entre las operaciones estadísticas de campo y la investigación de las universidades e institutos. Coincidimos con Dillman en que es difícil lograr el cambio.

La aplicación de una nueva metodología debe contar con el apoyo no sólo de los directores, sino también de un diálogo entre los investigadores que formulan métodos innovadores y quienes trabajan con las encuestas que deben mejorarse. Los métodos que proponemos se han formulado al mismo tiempo que analizamos problemas de registro con los responsables de los productos de registros. Hemos dedicado mucho tiempo en seminarios y círculos de estudios a promover ideas y métodos novedosos. Este trabajo de aplicación práctica será necesario en todas las oficinas de estadística donde se lleven a cabo nuevas estadísticas basadas en registros.

Se requiere un nuevo enfoque

Es necesario adoptar un nuevo enfoque para el manejo de datos administrativos:

– El prejuicio sobre la mala calidad de los datos administrativos no tiene razón de ser. Si comparamos la calidad de las enormes cantidades de datos administrativos que las autoridades hacendarias recopilan por medio de los formularios fiscales de personas y empresas con la calidad de los mismos datos captados por una oficina de estadística, habremos de admitir que los datos recopilados por las autoridades hacendarias son de mejor calidad. Scheuren y Petska (1993) opinan que “los datos detallados sobre ingresos y gastos que aparecen en la declaración de impuestos suelen considerarse más confiables que datos similares obtenidos en una encuesta”.

– Por otro lado, los datos administrativos no se deben usar tal como están, sino que se deben procesar para fines estadísticos. La parte más importante de este procesamiento es la integración de muchas fuentes.

Por todo esto, es necesario adoptar un nuevo enfoque respecto a los registros y las ciencias estadísticas. Algunos estadísticos dicen que no hay problemas metodológicos propios de las estadísticas basadas en registros y que no hay diferencia respecto a los censos comunes. Este malentendido se debe a que no se han percatado de los problemas metodológicos exclusivos de las encuestas basadas en registros. La fase de integración de las encuestas basadas en registros y los métodos utilizados determinan en buena medida la calidad y no se parecen a los métodos aplicados en los censos.

Por consiguiente, el desarrollo de la estadística basada en registros se debe reconocer como un importante campo de las ciencias estadísticas.

El sistema de registros

Se usan datos administrativos de muchas fuentes para crear un sistema coordinado de registros estadísticos. Este sistema puede servir para producir estadísticas basadas en registros, generar nuevos registros y establecer marcos para encuestas por muestreo o censos. Si el sistema de registros se creó de manera adecuada, será un factor importante para fomentar la consistencia y la coherencia entre todas las encuestas realizadas por una oficina de estadística. Incluso los países que apenas comienzan a producir estadísticas basadas en registros se verán beneficiados si coordinan los registros que tienen en un sistema.

Cabe señalar que nuestra propuesta es crear *un* solo sistema para todas las encuestas. Si se crean diferentes subsistemas (por ejemplo, uno para estadísticas sociales y otro para estadísticas económicas), será difícil combinar los respectivos datos. La productividad por rama industrial se puede estimar dividiendo:

– el valor agregado por rama industrial, de acuerdo con una encuesta basada en registros —en este caso el Registro Empresarial—, entre
– las horas trabajadas por rama industrial, de acuerdo con la Encuesta de Fuerza Laboral, que es una encuesta por muestreo basada en el Registro Población.

Para obtener un producto de buena calidad en este ejemplo, el sistema de registros debe ser *un solo* sistema coordinado y en los métodos de estimación se debe considerar que la rama industrial es una variable de valores múltiples. Los errores de integración, que en este caso son errores de agregación, pueden dar origen a inconsistencias en la rama industrial del Registro Empresarial y la Encuesta de Fuerza Laboral.

La principal ventaja de las estadísticas basadas en registros es el buen nivel de cobertura y consistencia que se obtiene si el sistema de registros se crea de acuerdo con los principios planteados en los capítulos 5 y 6. Todas las encuestas basadas en el sistema de registros se pueden beneficiar de ello y las Cuentas Nacionales recabarán datos más consistentes sin subcobertura.

Errores de los marcos

Cochran (1963) dedica doce de los trece capítulos de su libro a los errores de muestreo. En el último capítulo, el autor menciona los errores de medición y la no respuesta. En las últimas décadas, se ha trabajado mucho en errores no muestrales. Hoy se considera que los problemas de medición y de no respuesta son centrales en la metodología de las encuestas.

Por su parte, Särndal y Lundström (2005) dedican trece de los catorce capítulos de su libro a los problemas de no respuesta y mencionan los errores de marco en el último capítulo. Hoy no existen métodos establecidos para manejar los errores de marco; pensamos que se ha soslayado este tipo de error no muestral, que puede llegar a ser sustancial. Es necesario avanzar en este campo y estos errores sólo se pueden reducir mediante métodos estadísticos para el aprovechamiento de registros. Si aprendemos a generar registros con una buena cobertura, todas las encuestas que usen estos registros se beneficiarán de ello.

Lo primero es identificar los errores de marco. En una oficina de estadística, donde la teoría de las encuestas por muestreo es el paradigma dominante, los registros se usan para producir marcos y después se captan los datos. Normalmente nunca se conocerá la calidad de la población marco, sino que se crearán nuevos marcos seguidos de una nueva ronda de captación de datos.

En una oficina de estadística, los responsables de, digamos, una encuesta de empresas que quieren usar datos administrativos deben seguir exactamente el mismo procedimiento, sólo que en vez de enviar cuestionarios a las empresas seleccionadas como muestra, usarán datos administrativos para la muestra. De este modo, la carga de la respuesta y los costos

disminuirán, pero los errores de los marcos seguirán siendo los mismos. Si se da este uso restrictivo a los datos administrativos, no se habrá aprovechado la principal cualidad de las fuentes administrativas: la capacidad de lograr una buena cobertura.

Si los registros se usan tanto para establecer marcos como para crear poblaciones por año calendario, será posible cobrar conciencia de los errores de los marcos. Las estimaciones preliminares para las encuestas por muestreo basadas en marcos también se pueden corregir con información del registro por año calendario, y los métodos aplicados para crear marcos también se pueden mejorar para reducir los errores de marco. La diferencia entre un marco de empresas basado en información preliminar y un registro empresarial por año calendario basado en todas las fuentes administrativas disponibles puede ser de alrededor de 10 por ciento de la facturación total, de acuerdo con nuestra experiencia en el Registro Empresarial sueco.

¿Qué más se necesita?

Además de mejorar los registros actuales, los funcionarios de la oficina de estadística deben descubrir constantemente *nuevas fuentes administrativas* que puedan servir para generar nuevos registros estadísticos y productos. Los nuevos tipos de registros y bases de datos que surgen *fuera del sector público* también pueden ser fuentes relevantes en el futuro. Las personas y las empresas empiezan a dejar numerosos expedientes electrónicos que las empresas privadas almacenan en bases de datos. Este tipo de registros no proviene de poblaciones relevantes definidas; la definición de una base de datos de una empresa está determinada por los contactos de la empresa con sus clientes, proveedores, etc. Para investigar la posición de una empresa en el mercado, resulta útil poder combinar los datos de la base de datos de la empresa con datos referentes a la población de registro relevante. En el futuro ésta podría ser una importante tarea para las oficinas nacionales de estadística: *crear poblaciones de registro relevantes* que satisfagan las necesidades de diferentes usuarios.

La capacidad de estructurar bases de datos para propósitos estadísticos y analizar los datos tomados de los sistemas administrativos de manera que tenga un significado estadístico será necesaria en muchos ámbitos nuevos. En consecuencia, se requieren habilidades estadísticas para el aprovechamiento de registros incluso más allá de los organismos gubernamentales a cargo de la realización de encuestas. Las *universidades e instituciones de educación superior* deben contar con programas de investigación y enseñanza sobre encuestas basadas en registros en relación tanto con la sociedad como con el sector empresarial.

Referencias

Bayard, K., Klimek, S. (2003): *Creating a Historical Bridge for Manufacturing Between the Standard Industrial Classification System and the North American Industry Classification System.* Actas de 2003 de la American Statistical Association, sección de estadísticas de negocios y económicas.

Bethlehem, J., Hoogland, J., Schouten, B. (2006): "From Surveys to Registers". Ponencia presentada en el cuarto Seminario sobre Estrategias para Estadísticas Sociales y Espaciales, Oslo, Noruega, 27-28 de febrero de 2006.

Biemer, P., Lyberg, L. (2003): *Introduction to Survey Quality*. John Wiley & Sons, Ltd.

Cochran, W. G. (1963): *Sampling Techniques*. 2a. edición, John Wiley & Sons, Ltd.

Deville, J., Särndal, C-E. (1992): "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, vol. 87, pp. 376-382.

de Waal, T., Quere, R. (2003): "A Fast and Simple Algorithm for Automatic Editing of Mixed Data". *Journal of Official Statistics*, vol. 19, núm. 4, pp. 383-404.

Dillman, D. (1996): "Why Innovation Is Difficult in Government Surveys". *Journal of Official Statistics*, vol. 12, núm. 2, pp. 113-124.

Eltinge, J., Kozlow, R., Luery, D. (2003): "Imputation in Three Federal Statistical Agencies". Artículo preparado para el Federal Economic Statistics Advisory Committee, octubre de 2003.

Eurostat (1997): *Proceedings of the Seminar on the Use of Administrative Sources for Statistical Purpose.* Luxemburgo, 15-16 de enero de 1997.

Federal Committee on Statistical Methodology (2005): *Report on Statistical Disclosure Limitation Methodology.* Statistical Policy Working Paper 22, 2a. versión.

Fellegi, I.P., Holt, D. (1976): "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, vol. 71, pp. 17-35.

Granquist, L., Kovar, J. (1997): "Editing of Survey Data: How Much Is Enough?". Lyberg *et al*. (compiladores), *Survey Measurement and Process Quality*, John Wiley & Sons, Ltd.

Greijer, Å. (1995): *Overcoverage of Foreign Born in the Population Register Estimated with the Labour Force Survey* (en sueco). Metodrapport från BoR-avdelningen 1995, vol. 3, Statistics Sweden.

Greijer, Å. (1996): *Overcoverage in the Population Register: An Analysis of Returned Mail* (en sueco). Metodrapport från BoR-avdelningen 1996, vol. 7, Statistics Sweden.

Greijer, Å. (1997a): *Overcoverage in the Population Register Estimated with the Income register* (en sueco). Metodrapport från BoR-avdelningen 1997, vol. 11, Statistics Sweden.

Greijer, Å. (1997b): *Overcoverage in the Population Register Estimated with the Labour Force Survey* (en sueco). Metodrapport från BoR-avdelningen 1997, vol. 12, Statistics Sweden.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., Tourangeau, R. (2004): *Survey Methodology*. John Wiley & Sons, Ltd.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Henell, S., Longhurst, J., Mitchell, B. (2003): *Small Area Estimation – Project Report.* Serie Model-Based Small Area Estimation, núm. 2, Office for National Statistics.

Holt, D. (2001): *Comment to Platek and Särndal.* Journal of Official Statistics, vol. 17, núm. 1, pp. 55-61.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., Snijders, V. (2003): *Estimating Consistent Table Sets: Position Paper on Repeated Weighting.* Statistics Netherlands, Documento de discusion 03005,2003.

Hörngren, J. (1992): *The Use of Registers as Auxiliary Information in the Swedish Labour Force Survey.* Informe de investigación y desarrollo, 1992:13, Statistics Sweden.

Johansson, D. (1997): *The Number and the Size Distribution of Firms in Sweden and Other European Countries.* IUI Documento de trabajo núm. 483, Research Institute of Industrial Economics, Estocolmo, Suecia.

Johansson, D. (2001): *The Dynamics of Firm and Industry Growth: the Swedish Computing and Communications Industry*. Royal Institute of Technology, Estocolmo, Suecia, TRITA-IEO R 2001:05.

Kardaun, J. W. P. F., Loeve, J. A. (2005): *Longitudinal Analysis in Statistical Offices.* Statistics Netherlands, Documento de discusión 05010.

Laan, P. van der (2002): *Creating a Social Statistics Database in the Netherlands: Progress and Priorities.* Artículo presentado en el segundo Seminario sobre Estrategias para Estadísticas Sociales y Espaciales, Copenhague, Dinamarca, 5 de septiembre de 2002.

Lindström, H. L. (1999): *Quality Assurance in Registers Based on Administrative Sources For Statistical Purposes* (en sueco). Rapport från Registerprojektet, Statistics Sweden.

Nanopoulos, P. (2001): *Comment to Platek and Särndal.* Journal of Official Statistics, vol. 17, núm. 1, pp. 77-86.

Pannekoek, J., de Waal, T. (2005): "Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project". *Journal of Official Statistics*, vol. 21, núm. 2, pp. 257-286.

Platek, R., Särndal, C-E. (2001): "Can a Statistician Deliver?". *Journal of Official Statistics*, vol. 17, núm. 1, pp. 1-20.

Statistics Sweden (2001): *The Future Development of the Swedish Register System: Final Report of the Register Project and Decision of the Director General.* Informe de investigación y desarrollo, Statistics Sweden, 2001:1.

Statistics Sweden (2001a): *Quality Concepts for Official Statistics.* MIS 2001:1, Statistics Sweden.

Scheuren, F. (1999): *Administrative Records and Census Taking.* Survey Methodology, vol. 25, núm. 2, pp. 151-160.

Scheuren, F., Petska, T. (1993): "Turning Administrative Systems into Information Systems". *Journal of Official Statistics*, vol. 9 núm. 1, pp. 109-119.

Selander, R., Svensson, J., Wallgren, A., Wallgren, B. (1998): *Administrative Registers in an Efficient Statistical system: New Possibilities for Agricultural Statistics? How Should We Use IACS Data?* Statistics Sweden y Eurostat.

Statistics Canada (2003): *Statistics Canada Quality Guidelines.* 4a. edición, octubre 2003.

Statistics Canada (2003): *The Integrated Approach to Economic Surveys in Canada.* UNECE, octava reunión del Grupo de Expertos en Cuentas Nacionales, abril 2006.

Statistics Denmark (1995): *Statistics on Persons in Denmark: a Register-based Statistical System.* Eurostat.

Statistics Finland (2004): *Use of Registers and Administrative Data Sources for Statistical Purposes: Best Practices of Statistics Finland.*

Statistics Netherlands (2000): *Special Issue: Integrating Administrative Registers and Household Surveys.* Netherlands Official Statistics, vol. 15.

Statistics Netherlands (2004): *The Dutch Virtual Census of 2001: Analysis and Methodology.*

Särndal, C-E., Lundström, S. (2005): *Estimation in Surveys with Nonresponse.* John Wiley & Sons, Ltd.

UNECE (1998): *Recommendations for the 2000 Censuses of Population and Housing in the ECE region.* Naciones Unidas.

U.S. Census Bureau (2003): *U.S. Census Bureau Strategic Plan FY 2004-2008.* Septiembre de 2003.

Wallgren, A., Wallgren, B., Perssson, R., Jorner, U., Haaland, J-A. (1996): *Graphing Statistics & Data: Creating Better Charts.* SAGE Publications.

Wallgren, A., Wallgren, B. (1998*): Linking a System of Time Series: Adjusted Estimates for the Labour Force Surveys 1987-1992* (en sueco). Bakgrundsfakta till Arbetsmarknads- och Utbildningsstatistiken 1998:2, Statistics Sweden.

Wallgren, A., Wallgren, B. (1999): *Administrative Registers in an Efficient Statistical System: How can we Use Multiple Administrative Sources?* Statistics Sweden y Eurostat.

Wallgren, A., Wallgren, B. (2002): "How to Achieve Good Quality of Register-based Statistics?". Ponencia presentada en el segundo Seminario sobre Estrategias para Estadísticas Sociales y Espaciales, Copenhague, Dinamarca, septiembre de 2002.

Wilén, A., Johannesson, I. (2002): *A New Total Population Register System: More Possibilities and Better Quality.* Bakgrundsfakta till Befolknings- och Välfärdsstatistik 2002, vol. 2, Statistics Sweden.

Willenborg, L., de Waal, T. (2001): *Elements Of Statistical Disclosure Control.* Springer.

Glosario

A lo largo del libro hemos introducido una serie de términos y principios con objeto de facilitar la elaboración de la metodología estadística para el aprovechamiento de registros. El propósito de este glosario es ofrecer un panorama general de esa terminología.

A

ajuste por no respuesta *(nonresponse adjustment)*	Método de estimación que tiene por objeto minimizar el error por no respuesta o los efectos de los valores perdidos.
anonimizado *(anonymised)*	Una matriz de datos anonimizados se crea eliminando todas las variables de identificación en un registro.

B

base de datos *(database)*	Las bases de datos que son usadas para almacenar microdatos estadísticos consisten en una serie de tablas de bases de datos relacionadas.
base de datos de clasificaciones *(classification database)*	Metadatos establecidos de manera formal para el manejo de todos los códigos aplicables a los estándares y clasificaciones estadísticas relevantes, así como las claves de conexión entre las diferentes versiones de estos estándares.
base de datos de definiciones *(definitions database)*	Metadatos establecidos de manera formal para manejar grandes cantidades de definiciones de variables.

C

calendario de eventos *(events calendar)*	Los cambios o eventos que influyen en el sistema de registros se deben documentar. Un calendario de eventos es un sistema informático con metadatos formalizados en el que es posible buscar información sobre eventos en relación con el tiempo, el registro y la variable.
calibración *(calibration)*	Método de estimación basado en ponderadores, donde los ponderadores se ajustan para minimizar los efectos de diferentes tipos de errores.
calidad de corte transversal *(cross-sectional quality)*	Se refiere a la calidad de la información de corte transversal y la posibilidad de comparar los datos del registro para un momento del tiempo o periodo específico.

Register-based Statistics – Administrative Data for Statistical Purposes A. Wallgren y B. Wallgren

calidad de las series de tiempo *(time series quality)*	Se refiere a calidad de la serie de tiempo de una variable y las comparaciones que se pueden hacer a lo largo del tiempo con los valores de las variables en el nivel agregado.
calidad longitudinal *(longitudinal quality)*	Se refiere a la calidad de los datos longitudinales y las comparaciones que se pueden hacer a lo largo del tiempo en variables en el *nivel micro*.
censo *(census)*	Encuesta con recopilación de datos propia en la que se recaban mediciones de las variables de interés para todos los objetos de la población. Los registros basados en censos se incluyen en el sistema de registro.
clasificación *(classification)*	Ejemplos de clasificaciones son: la clasificación industrial, el clasificador de productos, la clasificación de programas educación, la clasificación de ocupaciones, etc., las cuales representan estándares estadísticos relevantes. Están basadas en recomendaciones internacionales y son importantes en lo que respecta a la posibilidad de homologar la forma de ordenar o agrupar el contenido de los resultados de las encuestas, tanto en encuestas basadas en registros como en encuestas por muestreo o censos. Las fuentes administrativas contienen datos que pueden ayudar a la adopción de las clasificaciones, que son útiles para crear variables dentro del sistema de registro.
clave externa *(foreign key)*	Las claves externas se usan para describir las relaciones entre objetos estadísticos diferentes. Véase *variable de referencia*.
clave primaria *(primary key)*	En informática, término para la variable de identificación de un registro dentro de la base de datos.
codificación *(coding)*	Los datos en forma de texto se pueden usar para crear variables estadísticas útiles para la medición y seguimiento de fenómenos sociales y económicos. Como parte del proceso de codificación, la información en forma de texto no estructurado se transforma en valores estructurados de variables.
coherencia *(coherence)*	Si las estimaciones de diferentes encuestas se pueden utilizar conjuntamente, hay coherencia entre ellas. Por ejemplo, para que un cociente tenga significado, el numerador y el denominador deben ser coherentes.
cohorte *(cohort)*	Grupo de objetos originado a partir de un periodo específico. Los registros longitudinales se pueden analizar dando seguimiento a diferentes cohortes a lo largo del tiempo.
conjunto de objetos (*object set*)	Concepto usado cuando describimos un registro sin referirnos a una encuesta específica. Sin embargo, cada registro estadístico se crea para uno o varios usos o encuestas principales. Por consiguiente, es común que el conjunto de objetos del registro concuerde con la población de la encuesta principal.
conjunto de valores *(value set)*	El conjunto de valores que puede asumir una variable, o se puede pensar que asume, para cualquier objeto.
consistencia *(consistency)*	Se considera que dos encuestas son consistentes si están relacionadas con la misma población y si los objetos tienen los mismos valores para variables comunes, con las mismas estimaciones.

D

datos agregados *(aggregated data)*	Datos relacionados con sumas de datos, valores resumen de los datos, macrodatos, que hacen referencia a grupos de objetos y normalmente se presentan como contenido en tablas estadísticos o tabulados.
documentación *(documentation)*	Trabajo comprendido en la creación de metadatos. Véase *metadatos*.

E

emparejamiento *(matching)*	Comparación de los vínculos entre dos o más registros. El resultado es una coincidencia o una no coincidencia (no emparejamiento). También se le conoce como emparejamiento *exacto*.
encuesta basada en registros *(register-based survey)*	Encuesta estadística sin un levantamiento de datos propio. En vez de ello, se usan registros administrativos o estadísticos.
encuesta de mantenimiento de registros *(register maintenance survey)*	Cuestionario enviado a los objetos de un registro base cuando faltan datos sobre ellos o se considera que la información del objeto es obsoleta.
encuesta por muestreo *(sample survey)*	Encuesta con levantamiento de datos propio, en la que se recolectan valores de las variables de interés en una muestra aleatoria de los objetos de la población.
error de agregación *(aggregation error)*	Error que puede presentarse cuando agregamos los valores de las variables con base en la relación *muchos objetos a un objeto* si la variable es cualitativa. Los errores de agregación pueden deberse a tres razones: ocurrencia múltiple de objetos en un registro, relaciones *muchos a uno* y variables con valores múltiples. Además de arrojar estimaciones erróneas, los errores de agregación pueden originar inconsistencias entre las estimaciones para diferentes registros.
error de clasificación *(classification error)*	El *error de clasificación bruto* estimado es una estimación de la proporción de clasificaciones incorrectas en todo el registro, mientras que la estimación del *error de clasificación neto* es una estimación del error sistemático.
error de cobertura *(coverage error)*	Diferencias entre la población objetivo y la población del registro. Véase *sobrecobertura* y *subcobertura*.
error de emparejamiento *(matching error)*	Coincidencias inexistentes o falsas cuando se efectúa el emparejamiento.
error de imputación *(imputation error)*	Valores imputados que difieren del valor correcto, el cual es desconocido.
error de integración *(integration error)*	Los tipos de errores originados en la fase de integración deben recibir el nombre de errores de integración. Esta categoría incluye los errores de cobertura, los errores de no emparejamiento, los valores perdidos y los errores de agregación.
error de modelo *(model error)*	Se pueden crear variables derivadas usando un modelo. Si los valores del modelo difieren de los valores correctos, se trata de un error de modelo.
error de relevancia (*relevance error*)	Si las definiciones de una población o una variable no son adecuadas para los objetivos de la encuesta, los resultados de la encuesta estarán sujetos a errores de relevancia.
estadísticas basadas en registros *(register-based statistics)*	Estadísticas producidas por medio de encuestas basadas en registros.

estándar estadístico (*standard*) — Véase *clasificación*.

estimación *(estimation)* — A partir de un registro estadístico existente, se pueden generar y publicar cuadros con estimaciones. Por lo general, el término *estimación* se usa en las encuestas por muestreo, pero también debe usarse en las estadísticas basadas en registros. Asimismo, es importante distinguir entre los valores reales y los valores estimados que produce el registro.

Dentro de las estadísticas basadas en registros, el método para crear un registro es el método de estimación *fundamental* pues determina qué estimaciones se pueden hacer. Para corregir problemas con variables de valores múltiples, sobrecobertura, valores perdidos y cambios de nivel en las series de tiempo, se pueden usar métodos de estimación *suplementarios*.

evento demográfico *(demographic event)* — Los objetos nacen, cambian de ubicación, se modifican o dejan de existir. Estos tipos de eventos reciben el nombre de eventos demográficos porque modifican la población.

I

imputación *(imputation)* — Los valores perdidos se remplazan por valores imputados, que se calculan usando una distribución de probabilidades o un modelo determinístico.

integración *(integration)* — En el trabajo de creación de un registro predomina la fase de integración, donde se combinan datos de muchas fuentes. Durante esta fase se crean la población del registro y los objetos derivados, se importan las variables de distintas fuentes y se generan variables derivadas.

instancia de objeto *(object instance)* — Término correspondiente a *objeto* en la modelación conceptual de un sistema informático. Esta terminología se refiere a una población, por ejemplo, de individuos cuyo tipo de objeto es el de persona, donde cada individuo de la población es una *instancia de objeto*. El término *objeto* se usa a menudo como sinónimo de *instancia de objeto*.

M

macrodatos *(macrodata)* — Datos agregados para grupos de objetos. Normalmente se presentan como contenido de cuadros estadísticos o tabulados.

macro-metadatos *(macro metadata)* — Metadatos para macrodatos. Describen el contenido de los cuadros estadísticos o de los tabulados, son los metadatos de los valores agregados de las variables.

matriz de datos *(data matrix)* — Datos estadísticos y microdatos ordenados de forma tal que las columnas contienen valores de las variables observadas y los renglones contienen observaciones de los objetos.

metadatos *(metadata)* — Información necesaria para usar e interpretar las estadísticas. Describe los datos mediante definiciones de las poblaciones, objetos, variables, metodología y calidad.

métodos de estimación complementarios *(supplementary estimation methods)* — Dentro de la estadística basada en registros, métodos de estimación que usan un registro estadístico junto con ponderadores o imputaciones.

métodos de estimación fundamentales *(fundamental estimation methods)*
Métodos para crear registros estadísticos dentro de la estadística basada en registros, es decir, cómo crear la población de registro y las variables en el registro.

microdatos *(microdata)*
Datos sobre objetos individuales.

micro-metadatos *(micro metadata)*
Metadatos para microdatos. Describen el contenido en las matrices de datos que contienen microdatos.

modelo conceptual *(conceptual model)*
Un modelo conceptual describe los tipos de objeto del sistema y sus relaciones independientemente de sus implementaciones físicas y técnicas.

N

no emparejamiento *(mismatch)*
Véase *emparejamiento*.

no respuesta parcial *(item nonresponse)*
Pueden faltar ciertos valores de variables para los objetos en el registro. Véase *valores perdidos*.

O

objeto *(object)*
Una población, registro o matriz de datos está constituida por varios objetos (elementos o unidades estadísticas).

objeto combinado *(combination object)*
Método de estimación para variables de valores múltiples en el que cada combinación de objetos y valores de las variables de valores múltiples corresponde a un objeto combinado en una nueva matriz de datos donde se usan ponderadores para la estimación.

objeto derivado *(derived object)*
Un objeto derivado se crea mediante un proceso de registro que usa información disponible sobre relaciones. Por ejemplo, las personas registradas en la misma vivienda constituyen hogares.

objeto relacional (*relational object*)
Una relación entre dos objetos puede considerarse un objeto relacional. Es necesario objetivar una relación cuando hay variables que deben vincularse con combinaciones de objetos y no con un objeto individual.

observación *(observation)*
Todos los valores obtenidos de las diferentes variables para un objeto específico. También recibe el nombre de entrada.

P

población *(population)*
La definición de población debe mostrar qué objetos están incluidos en la población. Se debe especificar el tipo de objeto. Siempre se debe incluir una referencia temporal y una delimitación geográfica. A su vez, en la delimitación geográfica se debe especificar la relación que existe entre el objeto o la unidad estadística y el área geográfica.

población de registro *(register population)*
Conjunto de objetos en el registro creado para la encuesta de que se trate, es decir, la población encuestada. El término correspondiente para censos y encuestas por muestreo es *población marco*.

población marco *(frame population)*
Conjunto de objetos definido por el marco de estudio y el proceso de establecimiento de dicho marco. Este término sólo se debe usar para censos y encuestas por muestreo. El término correspondiente para las encuestas basadas en registros es población de registro.

población estandarizada *(standardised population)*
: Todas las poblaciones deben crearse usando registros base. Un uso importante de los registros base es la creación de poblaciones estandarizadas, que se usan, a su vez, para crear poblaciones de registro en los otros registros del sistema.

  Las poblaciones estandarizadas deben ser de buena calidad en lo que respecta a la cobertura y las variables de expansión importantes. Por lo tanto, se crean luego del periodo o momento al que se refieren cuando se ha notificado toda la parte administrativa. Después, si otros registros usan exclusivamente estas poblaciones de registro estandarizadas, las estadísticas basadas en registros serán consistentes en relación con los conjuntos de objetos.

ponderador *(weight)*
: En las matrices de datos de las muestras, cada observación en la muestra puede representar muchas observaciones en la población. Una parte importante del trabajo en las encuestas por muestreo consiste en calcular los ponderadores o los coeficientes de ajuste.

R

recopilación de datos integrada *(integrated data collection)*
: Una autoridad administrativa recopila información que no se usa para fines administrativos, pero son datos solicitados por la autoridad estadística. Como ejemplo tenemos los datos de los estados de resultados, en cuyo caso el Consejo Nacional Tributario de Suecia también recopila cierta información para la Oficina de Estadística de Suecia, la cual no es relevante para el trabajo de la autoridad tributaria.

referencia temporal *(time reference)*
: Variable que muestra el momento en que ocurre un evento que afecta los objetos o las actualizaciones del registro.

registro (*register*)
: Un registro tiene la finalidad de ser una lista completa de los objetos en un grupo de objetos o una población específicos. Sin embargo, pueden faltar datos sobre algunos registros a causa de deficiencias en la calidad. Los datos sobre la identidad de un objeto también deben estar disponibles, de modo que el registro se pueda actualizar y ampliar con nuevos valores de las variables para cada objeto. Por ende, las características más importantes de un registro son que el listado está completo y se conocen las identidades de los objetos incluidos en el listado.

registro administrativo *(administrative register)*
: Registro usado para fines administrativos en un sistema de información administrativa. Contendrá todos los objetos por administrar, sus objetos serán identificables y sus variables se usarán para propósitos administrativos.

registro de inventario actual *(current stock register)*
: Versión del registro puesta al día con toda la información disponible para relacionar los objetos actualmente activos o vivientes. Se usa como población de estudio.

registro base *(base register)*
: Registro estadístico de gran importancia para todo el sistema de registro. Debe definir tipos de objeto importantes, conjuntos de objetos importantes o poblaciones estandarizadas, y contiene vínculos con objetos en otros registros base.

registro de eventos *(events register)*
: Un registro de eventos para un periodo específico es un registro que contiene información sobre los eventos demográficos que han ocurrido durante el periodo. Se crea un registro por cada tipo de ocurrencia.

registro de fin de año *(end of year register)*
: Versión del registro referida al 31 de diciembre o al 1 de enero. Es útil para las estadísticas de población anuales y constituye un ejemplo de registro que se refiere a un momento específico en el tiempo.

registro de fuentes *(source register)*
Término referido tanto a las fuentes administrativas como a los registros de la Oficina de Estadística de Suecia que se usan para crear un nuevo registro.

registro en un momento específico *(register at a specific point in time)*
El registro referido a un momento específico —como el cambio de año— es la versión del registro actualizada para describir el conjunto de objetos en dicho momento. Esta actualización se lleva a cabo luego de ese momento, cuando se cuenta con información sobre todos los eventos hasta entonces. Se usa para las encuestas basadas en registros.

registro estadístico *(statistical register)*
Registro procesado para propósitos estadísticos. Se crean mediante el procesamiento de registros administrativos de modo que los conjuntos de objetos, los objetos y las variables satisfagan necesidades estadísticas.

registro estadístico primario *(primary statistical register)*
Un registro primario es un registro estadístico que se basa directamente en cuando menos una fuente administrativa. Los registros primarios están basados en fuentes administrativas, con las que se crea la parte principal de las variables estadísticas en el sistema de registro.

registro histórico *(historical register)*
Registro que contiene información sobre todos los eventos demográficos que han ocurrido en relación con determinado objeto. Por ejemplo, un objeto que tiene tres eventos se puede encontrar en tres renglones de la matriz de datos.

registro integrado *(integrated register)*
Registro estadístico que se creó *exclusivamente* combinando información que ya existía en los registros estadísticos del sistema.

registro longitudinal *(longitudinal register)*
Registro integrado donde se han compilado registros para diferentes momentos o periodos, de modo que es posible dar seguimiento al mismo objeto a lo largo del tiempo.

registro por año calendario *(calendar year register)*
Versión del registro que contiene todos los objetos que han existido en cualquier momento durante un año específico. Los objetos que se agregan o dejan de existir durante el año se deben incluir junto con la información sobre la fecha del evento.

relación *(relation)*
Entre diferentes objetos de una misma clase de objeto puede haber relaciones de diferentes tipos, por ejemplo, padre-hijo. También puede haber relaciones entre objetos de diferentes clases, por ejemplo, empleado-empleador. Cuando hay muchas relaciones importantes desde un punto de vista jurídico, los datos sobre relaciones a menudo se incluyen en el registro administrativo. Con estas relaciones se pueden crear los vínculos en el sistema de registro.

riesgo de revelación de datos confidenciales *(disclosure risk)*
Si es posible derivar características delicadas sobre personas o empresas a partir de una tabla estadística o tabulado, la publicación de ese tabulado supone un riesgo de revelación de datos confidenciales.

S

selección *(selection)*
Se crea un nuevo registro eligiendo ciertos objetos y variables de registros ya existentes. Diferenciamos entre *selección de objetos* y *selección de variables*.

sobrecobertura (*overcoverage*)
El registro contiene objetos que no pertenecen a la población objetivo.

subcobertura *(undercoverage)*
Objetos pertenecientes a la población objetivo no presentes en el registro.

T

tabla *(table)*
: Una tabla estadística o tabulado contiene datos agregados o datos a nivel macro y una tabla de base de datos contiene microdatos.

tabla de base de datos *(database table)*
: Cuando la Oficina de Estadística de Suecia cambió las computadoras centrales por servidores de bases de datos, se remplazaron antiguos términos como *archivo plano con registros y posiciones* por *tabla de base de datos con renglones y columnas*. Una matriz de datos se puede guardar en una tabla de base de datos.

tipo de objeto *(object type)*
: Tipo de objeto que constituye un registro o una población. Por ejemplo, persona, hogar, establecimiento, propiedad.

V

validación de la consistencia *(consistency editing)*
: Cuando se valida la consistencia en una encuesta basada en registros, se validan datos de distintas fuentes y los errores sospechados pueden ser producto de errores en las variables y en los objetos.

valores perdidos *(missing values)*
: Pueden presentarse una no respuesta parcial o valores perdidos en los registro a causa de fallas en el sistema administrativo, no emparejamiento o rechazo de ciertos valores de variables durante la validación.

variable (*variable*)
: Atributo medible de un objeto.

variable adjunta *(adjoined variable)*
: Variable derivada en un registro creado con variables de otro registro que tiene objetos *diferentes*. Los objetos del primer registro fuente pueden vincularse con objetos del segundo registro fuente en una relación *uno a uno* o en una relación *uno a muchos*.

variable agregada *(aggregated variable)*
: Variable derivada en un registro creado con variables de otro registro que tiene objetos *diferentes*. Los objetos del registro fuente se pueden vincular con los objetos del registro objetivo usando una relación *muchos a uno*. Es posible agregar valores, de manera que sea relevante para la encuesta y para los *micro-objetos* en el registro fuente vinculado con el *macro-objeto* respectivo en el registro objetivo.

variable auxiliar *(auxiliary variable)*
: Variable del registro usada durante la fase de estimación en una encuesta por muestreo.

variable de comunicación *(communication variable)*
: Las variables de comunicación, como nombre, domicilio y número telefónico, se usan cuando la Oficina de Estadística de Suecia necesita ponerse en contacto con un objeto respecto a un cuestionario o una entrevista.

variable de expansión *(spanning variable)*
: Variable que define las celdas en los cuadros estadísticos. En un cuadro unidireccional hay una variable de expansión, en un cuadro bidireccional dos variables de expansión, etcétera.

variable de flujo *(flow variable)*
: Variable que muestra valores o cantidades para diferentes periodos, por ejemplo, ingreso salarial durante un año.

variable de identificación *(identifying variable)*
: Variable utilizada para identificar claramente los objetos. El término correspondiente en informática es *clave primaria*.

variable de inventario *(stock variable)*
: Variable que muestra características en un momento específico, por ejemplo, la edad de una persona en un momento específico.

variable de referencia *(reference variable)*	Las variables de referencia (claves externas) se usan para describir las relaciones entre diferentes objetos. Cuando se emparejan registros que contienen datos sobre diferentes objetos, las variables de referencia producen coincidencias entre objetos relacionados.
variable de respuesta (*response variable*)	Por cada celda en un cuadro, se calculan medidas descriptivas de las *variables de respuesta*.
variable de valor único *(single-valued variable)*	Variable que sólo puede tener un valor por cada objeto observado.
variable de valores múltiples *(multi-valued variable)*	Variable que tiene más de un valor para al menos un objeto. Por ejemplo, la variable *ocupación* se puede describir mediante un valor para quienes tienen una sola ocupación, pero se debe describir con más de un valor para quienes tienen más de una ocupación.
variable de vinculación *(linkage variable)*	Un vínculo entre dos registros consiste en una o más variables de vinculación. Una variable de vinculación representa información que identifica un objeto específico o un grupo de objetos.
variable derivada *(derived variable)*	Nueva variable formada a partir de variables ya existentes. Las variables pueden derivarse mediante agrupamiento, clasificación, cálculos, adjunción o agregación. También distinguimos entre variables derivadas locales e importadas.
variable derivada localmente *(locally derived variable)*	Variable derivada que se ha formado en el registro relevante.
variable estandarizada *(standardised variable)*	Variable de significado fundamental que se usa en varios registros. Hay una responsabilidad clara por esta variable en cuanto a definiciones, asignación de nombres y documentación.
variable generadora de ponderadores *(weight-generating variable)*	Variable de registro usada para crear los ponderadores necesarios a fin de hacer estimaciones para variables de varios valores.
variable importada *(imported variable)*	Variable tomada de otro registro estadístico.
variable local *(local variable)*	Una variable que se forma por primera vez en el sistema de registro, dentro de cierto registro es una variable local para ese registro.
variable primaria *(primary variable)*	Una variable de un registro puede ser primaria o derivada. Las variables tomadas directamente de un registro administrativo reciben el nombre de *variables primarias locales*.
variable primaria local *(local primary variable)*	Variable primaria tomada directamente de un registro administrativo.
variable técnica *(technical variable)*	Variable para la administración interna de un registro.
variación aleatoria natural *(natural random variation)*	Las estimaciones basadas en registros pueden verse afectadas por la variación aleatoria natural, razón por la cual las diferencias y los cambios deben analizarse para tener una interpretación de los cambios.

vinculación de series de tiempo *(linking of time series)*	Método de estimación que tiene por objeto aumentar la comparabilidad a lo largo del tiempo.
vínculo *(link)*	Consiste en una o más variables de vinculación que identifican objetos. Si estas variables son de buena calidad y están presentes en varios registros, se pueden usar como vínculos cuando se emparejan varios registros. Un vínculo consiste en información que identifica objetos individuales.
vista *(view)*	Término propio de las bases de datos que se refiere a una presentación ajustada del contenido de los cuadros de las bases de datos.

Índice

Estadísticas basadas en registros
Aprovechamiento estadístico de datos administrativos

Anders Wallgren y Britt Wallgren
Oficina de Estadística de Suecia

Hay un interés creciente en el desarrollo de las encuestas basadas en registros, es decir, las encuestas que se apoyan en datos administrativos ya disponibles. Dado que se generan enormes cantidades de datos dentro de los diversos sistemas administrativos, existe la oportunidad de utilizarlos para el análisis estadístico sin incurrir en los costos que implica la captación de la información.

Las encuestas basadas en registros requieren una metodología propia, cuya formulación representa un reto importante para la ciencia estadística. En lugar de métodos para captar datos, son necesarios métodos para integrar datos de diferentes fuentes. ¿Cómo deben transformarse los datos administrativos para satisfacer las necesidades estadísticas?

Estadísticas basadas en registros ofrece a sus lectores una exposición detallada de los principios y prácticas de esta área cada vez más difundida de la estadística.

- Presenta una visión global de las estadísticas basadas en registros, en lo que respecta tanto a su teoría como a su aplicación avanzada.
- Utiliza ejemplos reales de la Oficina de Estadística de Suecia para ilustrar principios fundamentales.
- Propone una terminología sistemática, indispensable, para este campo de la estadística.
- Describe cómo crear registros estadísticos y una metodología para la integración de datos de múltiples fuentes como una herramienta fundamental para el futuro.
- Elabora métodos de estimación y conceptos relativos a la calidad para las encuestas basadas en registros.
- Analiza sistemas estadísticos integrados por muchos registros estadísticos y encuestas, destacando la importancia de la consistencia y la coherencia.

Estadísticas basadas en registros constituye una guía única para todos aquellos que laboran en una oficina de estadística. También es un valioso apoyo para los investigadores académicos y profesores en el campo de la estadística, así como para los estadísticos que trabajan con sistemas administrativos en instituciones gubernamentales y empresas.

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA MÉXICO
www.inegi.org.mx
ISBN 978-607-494-313-9