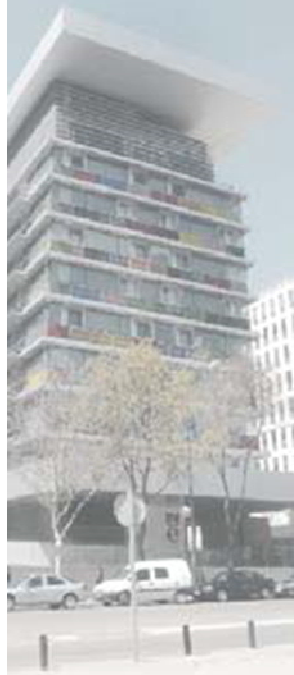




INSTITUTO NACIONAL DE ESTADISTICA



Secreto Estadístico

*Instituto Nacional de Estadística
Curso selectivo 2016*

Madrid, 29 de junio de 2016



Introducción

¿A qué nos dedicamos?

El INE se dedica sobre todo a hacer estadísticas a partir de la información obtenida de registros administrativos o de preguntar directamente al informante: personas físicas o empresas.

Esto nos lleva a acumular gran cantidad de información individualizada que puede ser un poco problemática de manejar.



Introducción

En general personas y empresas somos reacias a dar información sobre nosotros mismos “así porque sí”: no nos apetece que nuestros datos estén en manos de otro y se conozcan, y nos preocupa el uso que se le pueda dar a esa información. Esta realidad está reconocida en la sociedad: hay leyes (y otras cosas) que protegen el derecho a la intimidad.

Por otro lado, nuestro trabajo es publicar datos elaborados a partir de esa información individualizada. A veces hasta publicamos la información individualizada (con muchos matices que veremos más adelante)



Introducción

La idea que debe quedar clara desde el principio al hablar de confidencialidad estadística, sin haber definido todavía nada, es que **no podemos publicar los datos de cualquier manera.**

Por un lado, **no podemos publicar información con nombres y apellidos.** Por otro lado, la información que publiquemos **no debería revelar información “implícita”.**



Introducción

¿Qué es la confidencialidad estadística?

La **confidencialidad** o **secreto estadístico (SE)** puede definirse como la obligación de los servicios estadísticos del Estado de **proteger la información** obtenida para la elaboración de estadísticas frente a su revelación ilegal y usos ilícitos.

A destacar: no importa el origen de la información (obtenida directa o indirectamente)



Introducción

Fundamentos:

- 1- Legales: leyes nacionales y europeas obligan a preservar el Secreto Estadístico
- 2- Éticos: Principio Fundamental de las Estadísticas Oficiales nº 6 de NNUU; Código de Buenas Prácticas de las Estadísticas Europeas, principio nº 5
- 3- Prácticos: si el informante no confía en que sus datos van a estar bien custodiados, no va a colaborar en las encuestas



Introducción

¿Es entonces importante?

Es MUY importante. Aparte de ser obligatorio, la confianza que la ciudadanía tiene en la institución es esencial para que los informantes colaboren y respondan de forma veraz a las encuestas



Legislación

Legislación “directa” (sobre Estadística):

- 1-Ley de la Función Estadística Pública de 1989 (LFEP)
- 2-Reglamento Europeo 223/2009 de las Estadística Europeas, y su modificación 2015/759
- 3-Reglamento Europeo 557/2013 de acceso a datos confidenciales con fines científicos

Legislación “indirecta” (la Estadística no es el tema central):
Ley Orgánica de Protección de Datos de Carácter Personal 1999 (LOPD)



Legislación: Definiciones útiles

Datos personales (LFEP): los referentes a personas físicas o jurídicas que o bien permitan la identificación inmediata de los interesados, o bien conduzcan por su estructura, contenido o grado de desagregación a la identificación indirecta de los mismos

Datos de carácter personal (LOPD): Cualquier información concerniente a personas físicas identificadas o identificables



Legislación: Definiciones útiles

Identificador directo: variables tipo nombre y apellidos, razón social, DNI, CIF etc. que conducen a unidad a la que pertenecen los datos con un grado de certeza seguro o casi seguro.

Los datos con identificadores directos en esta presentación se denominan “datos identificados”

Identificador indirecto: variables que por sí mismas no conducen a la unidad a la que pertenecen los datos, pero cuya combinación y grado de desagregación sí puede llevar a identificar al informante: edad, profesión, municipio de residencia, rama de actividad etc.



Legislación: Definiciones útiles

Uso estadístico: utilizar la información para producir estadísticas con el fin de describir la realidad

Uso administrativo: uso de la información para tomar decisiones sobre una persona física o jurídica (conceder o no una beca, poner una multa, permitir una actividad en un local, facilitar la inscripción de los niños en el colegio, etc.)



Legislación

Ley de la Función Estadística Pública:

Capítulo III “Del Secreto Estadístico”, artículos 13-19



Ley de la Función Estadística Pública

Artículo 13:

1. Serán objeto de protección y quedaran amparados por el secreto estadístico los datos personales que obtengan los servicios estadísticos tanto directamente de los informantes como a través de fuentes administrativas.
2. Se entiende que son datos personales los referentes a personas físicas o jurídicas que o bien permitan la identificación inmediata de los interesados, o bien conduzcan por su estructura, contenido o grado de desagregación a la identificación indirecta de los mismos.
3. El secreto estadístico obliga a los servicios estadísticos a no difundir en ningún caso los datos personales cualquiera que sea su origen.



Ley de la Función Estadística Pública

Artículo 14:

- 1.El secreto estadístico será aplicado en las mismas condiciones establecidas en el presente capítulo frente a todas las Administraciones y organismos públicos, cualquiera que sea la naturaleza de éstos, salvo lo establecido en el artículo siguiente.
2. Queda prohibida la utilización para finalidades distintas de las estadísticas de los datos personales obtenidos directamente de los informantes por los servicios estadísticos.



Legislación

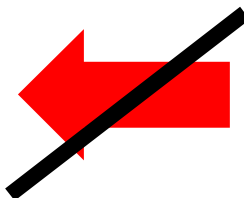
Datos administrativos



Datos estadísticos

Permitido. Deseable. Incentivado. La
tendencia actual

Datos administrativos



Datos estadísticos

Prohibidísimo. Penado



Ley de la Función Estadística Pública

Artículo 15:

1. La comunicación a efectos estadísticos entre las Administraciones y organismos públicos de los datos personales protegidos por el secreto estadístico solo será posible si se dan los siguientes requisitos, que habrán de ser comprobados por el servicio u órgano que los tenga en custodia:

- Que los servicios que reciban los datos desarrollen funciones fundamentalmente estadísticas y hayan sido regulados como tales antes de que los datos sean cedidos.
- Que el destino de los datos sea precisamente la elaboración de las estadísticas que dichos servicios tengan encomendadas.
- Que los servicios destinatarios de la información dispongan de los medios necesarios para preservar el secreto estadístico.



Ley de la Función Estadística Pública

Artículo 15 (cont.):

2. La comunicación a efectos no estadísticos entre las Administraciones y organismos públicos de la información que obra en los Registros públicos, no estará sujeta al secreto estadístico, sino a la legislación específica que en cada caso sea de aplicación.



Ley de la Función Estadística Pública

Artículo 16:

1. No quedarán amparados por el secreto estadístico los directorios que no contengan más datos que las simples relaciones de establecimientos, empresas, explotaciones u organismos de cualquier clase, en cuanto aludan a su denominación, emplazamiento, actividad y el intervalo de tamaño al que pertenece.
2. El dato sobre el intervalo de tamaño solo podrá difundirse si la unidad informante no manifiesta expresamente su disconformidad.



Ley de la Función Estadística Pública

Artículo 16 (cont.):

3. Los servicios estadísticos harán constar esta excepción a la preservación del secreto estadístico en los instrumentos de recogida de la información.

4. Los interesados tendrán derecho de acceso a los datos personales que figuren en los directorios estadísticos no amparados por el secreto y a obtener la rectificación de los errores que contengan.



Ley de la Función Estadística Pública

Artículo 16 (cont.):

5. Las normas de desarrollo de la presente Ley establecerán los requisitos necesarios para el ejercicio del derecho de acceso y rectificación a que se refiere el apartado anterior de este artículo, así como las condiciones que habrán de tenerse en cuenta en la difusión de los directorios no amparados por el secreto estadístico.



Ley de la Función Estadística Pública

Artículo 17:

1. Todo el personal estadístico tendrá la obligación de preservar el secreto estadístico.
2. A los efectos previstos en el párrafo anterior, se entiende por personal estadístico el dependiente de los servicios estadísticos a que aluden los títulos II y III de la presente Ley.



Ley de la Función Estadística Pública

Artículo 17 (cont.):

3. Quedarán también obligados por el deber de preservar el secreto estadístico cuantas personas, físicas o jurídicas, tengan conocimiento de datos amparados por aquel con ocasión de su participación con carácter eventual en cualquiera de la fases del proceso estadístico en virtud de contrato, acuerdo o convenio de cualquier género.

4. El deber de guardar el secreto estadístico se mantendrá aun después de que las personas obligadas a preservarlo concluyan sus actividades profesionales o su vinculación a los servicios estadísticos.



Ley de la Función Estadística Pública

Artículo 18:

1. Los datos que sirvan para la identificación inmediata de los informantes se destruirán cuando su conservación ya no sea necesaria para el desarrollo de las operaciones estadísticas.
2. En todo caso, los datos aludidos en el apartado anterior se guardarán bajo claves, precintos o depósitos especiales.



Ley de la Función Estadística Pública

Artículo 19:

1. La obligación de guardar el secreto estadístico se iniciará desde el momento en que se obtenga la información por él amparada.
2. La información a que se refiere el apartado anterior no podrá ser públicamente consultada sin que medie consentimiento expreso de los afectados o hasta que haya transcurrido un plazo de veinticinco años desde su muerte, si su fecha es conocida o, en otro caso, de cincuenta años a partir de la fecha de su obtención.



Ley de la Función Estadística Pública

Artículo 19 (cont.):

3. Excepcionalmente, y siempre que hubieran transcurrido, al menos, veinticinco años desde que se recibió la información por los servicios estadísticos, podrán ser facilitados datos protegidos por el secreto estadístico a quienes, en el marco del procedimiento que se determine reglamentariamente acrediten un legítimo interés.

4. En el caso de los datos relativos a personas jurídicas, las normas reglamentarias, atendidas las peculiaridades de cada encuesta, podrán disponer períodos menores de duración del secreto, nunca inferiores a quince años.



Secreto estadístico: Principales (que no únicas) implicaciones

- 1- No revelar nunca datos que permitan la identificación de las unidades estadísticas, cualquiera que sea su origen
- 2- Los datos recogidos con fines estadísticos solo pueden usarse con fines estadísticos
- 3- Todo el personal que participe en el proceso estadístico que tenga conocimiento de datos protegidos por SE tiene la obligación de preservarlo desde el momento en que tiene conocimiento de tales datos. Esta obligación no se extingue.



Legislación

+ El Reglamento Europeo 223/2009

Dice cosas muy similares, y contiene el siguiente artículo:

Artículo 23 (incompleto):

Acceso a datos confidenciales con fines científicos

La Comisión (Eurostat) o los INE u otras autoridades nacionales, en sus respectivas esferas de competencia, podrán conceder el acceso a datos confidenciales que solo permitan la identificación indirecta de unidades estadísticas a investigadores que lleven a cabo análisis estadísticos con fines científicos. Si los datos han sido transmitidos a la Comisión (Eurostat), se requerirá la aprobación del INE u otra autoridad nacional que proporcionó los datos.



Legislación

La Ley Orgánica de Protección de Datos de Carácter Personal (LOPD)

Contempla excepciones para la actividad estadística:

Artículo 2.3 c de la LOPD 1999:

Se registrarán por sus disposiciones específicas, y por lo especialmente previsto, en su caso, por esta Ley Orgánica los siguientes tratamientos de datos personales:

c) Los que sirvan a fines exclusivamente estadísticos, y estén amparados por la legislación estatal o autonómica sobre la función estadística pública



Legislación

LOPD: solo datos concernientes a personas FÍSICAS

Contempla derechos ARCO (acceso - rectificación - cancelación - oposición):

Acceso: ¿qué datos tienes míos?

Rectificación: los datos que tienes están mal, corrígelos

Cancelación: los datos que tienes son excesivos, bórralos

Oposición: el tratamiento que vas a dar a mis datos no me gusta, no tienes mi permiso

Allí donde la LFEP no llega, se aplica la LOPD



Protección de la información

A efectos prácticos: ¿qué es lo que tenemos que proteger?

El Secreto Estadístico en la práctica consiste en **evitar que un tercero obtenga información sobre cualquier informante con “nombre y apellidos”**

La protección del SE debe estar presente en **todas las fases de proceso estadístico**; no es cosa exclusiva de la difusión. Desde el momento en el que se diseña la encuesta se debe tener presente que se va a trabajar con datos confidenciales.



Protección de la información

Pensemos en cómo se produce una revelación indebida de datos:

Si los datos están identificados (con identificadores directos), para que se produzca una revelación indebida basta con que alguien no autorizado acceda a esos datos

Si los datos no están identificados, la revelación indebida se produce si la persona que consulta los datos logra deducir a qué informante corresponden los datos y obtiene con ello información de la que no disponía previamente (matizaremos esta afirmación más adelante)



Protección de la información

¿Qué puede pasar durante la recogida?

- Que un ajeno oiga las respuestas que da el informante, (se aplica en encuestas sobre temas sensibles como salud, hábitos sexuales, consumo de drogas etc.).
- Solución: hacer que el informante dé las respuestas codificadas, cuestionario autoadministrado
- Que el entrevistador se le caigan los cuestionarios y no se dé cuenta.
- Solución: hacer la recogida por medios digitales que encripten la información



Protección de la información

¿Qué puede pasar durante el procesamiento de la información?

- Que alguien (trabajador o externo) saque la información del INE.
- Solución: medidas de seguridad físicas, lógicas y administrativas: cortafuegos, control de acceso a los equipos informáticos, eliminación de los identificadores directos lo antes posible, almacenamiento de cuestionarios en armarios con llave, control de acceso al edificio, elaboración de normas de seguridad para los empleados etc.

Protección de la información

¿Qué puede pasar al publicar la información?

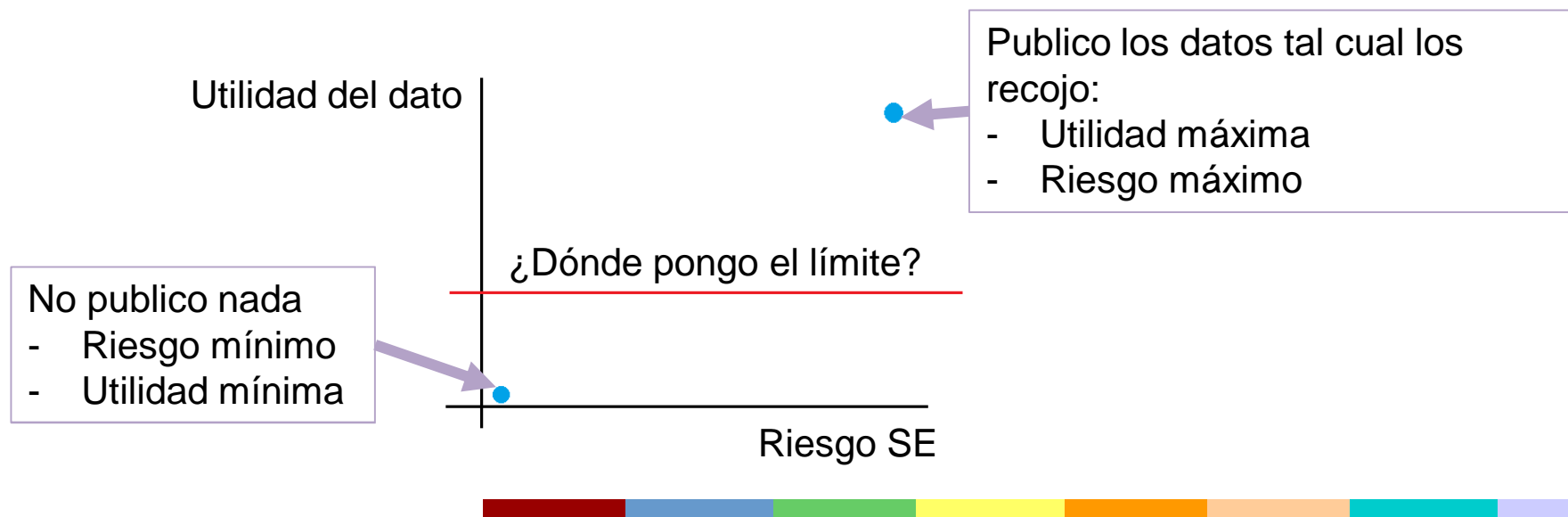
La publicación de la información es un poco especial. La protección del SE en las fases de recogida y procesamiento es “sencilla” en el siguiente sentido: el objetivo es evitar que un tercero logre acceder a los datos.

Pero cuando publicamos los datos los ponemos a propósito a disposición de terceros. Así que la protección del SE es más un problema “estadístico” que “físico”: debemos evitar que a partir de los datos publicados el usuario sea capaz de deducir a quién pertenecen (y obtener información extra)



Protección de la información

Este problema tiene más fondo del que podría parecer, porque uno de nuestros objetivos como productor de estadísticas públicas es que los datos sean lo más aprovechables posible. Pero **aumentar la utilidad de los datos choca con la obligación de preservar la confidencialidad.**



Protección de la información

Al publicar datos mínimamente útiles siempre se corre un riesgo en términos de SE. Por eso no tiene sentido adoptar políticas de evitar el riesgo, sino que se opta por políticas de gestión del riesgo.

La primera de las medidas: **nunca se publica, ni se facilita el acceso, a datos con identificadores directos**



Protección de la información

¿Cómo se produce la identificación al publicar de datos?

El usuario debe disponer de conocimientos previos sobre las unidades informantes, y cruza esa información con la información publicada.

Los conocimientos previos que tenga el usuario no los controlamos, pero la información que publicamos nosotros, sí.



Protección de la información

Ejemplo:

Info. Usada para identificar

Municipio	Edad	Estado civil	Ingresos mes
Villa Feliz (2000 habitantes)	29	Viuda	5000

Info. Extra que obtengo

Los habitantes del pueblo pueden identificar fácilmente a la persona y conocer su nivel de ingresos

Protección de la información

¿Qué facilita la identificación?

Si no hay identificadores directos (y nunca debería haberlos) las unidades en riesgo de ser identificadas son aquellas con características poco frecuentes en un ámbito geográfico o temporal dado: hablamos de unidades **raras**

La información muy detallada (tablas que cruzan muchas variables, variables con muchas categorías etc.) fragmenta mucho la población y puede facilitar la identificación



Protección de la información

A tener en cuenta al proteger la información:

No es lo mismo que los datos procedan de un censo / registro administrativo a que procedan de una muestra:

Censo: hay mucha certeza sobre que los datos de todo el mundo están ahí

Muestra: no hay tal certeza. Corolario: **la muestra debe ser confidencial**



Protección de la información

A tener en cuenta al proteger la información:

No es lo mismo que los datos sean sobre hogares o sobre empresas:

Las encuestas de tipo económico (empresas) son más difíciles de proteger: población “pequeña” y con unidades autorrepresentadas.



Protección de la información

A tener en cuenta al proteger la información:

No es lo mismo proteger microdatos que tablas de resultados: se aplican procedimientos distintos.

Microdatos: conjunto de los datos individuales de cada uno de los informantes

Campos

Nombre	DNI	Sexo	Edad	Estudios	Ocupación	Ingresos	Est. civil	Municipio
Pepe	12345678A	1	50	Licenc.	Profesor	1500	Casado	Madrid
Ana	87654321Z	6	32	Licenc.	Médico	2200	Soltera	Pinto

Registros



Protección de la información

Tabla: agregación de los microdatos

Nombre empresa	Actividad	Ingresos	Gastos de personal
Hotel 1	Hostelería	200.000	75.000
Pensión 1	Hostelería	83.000	22.000
...
Revista 1	Info. y com.	1.200.000	550.000
Periódico 2	Info. y com.	95.000	56.000



	Nº Empresas	Volumen negocio	Gastos de personal
Hostelería	283.868	58.229.416	18.450.239
Información y comunicaciones	48.268	81.819.008	18.059.555



Protección de la información

Distinguimos además entre tablas de magnitudes y de frecuencias:

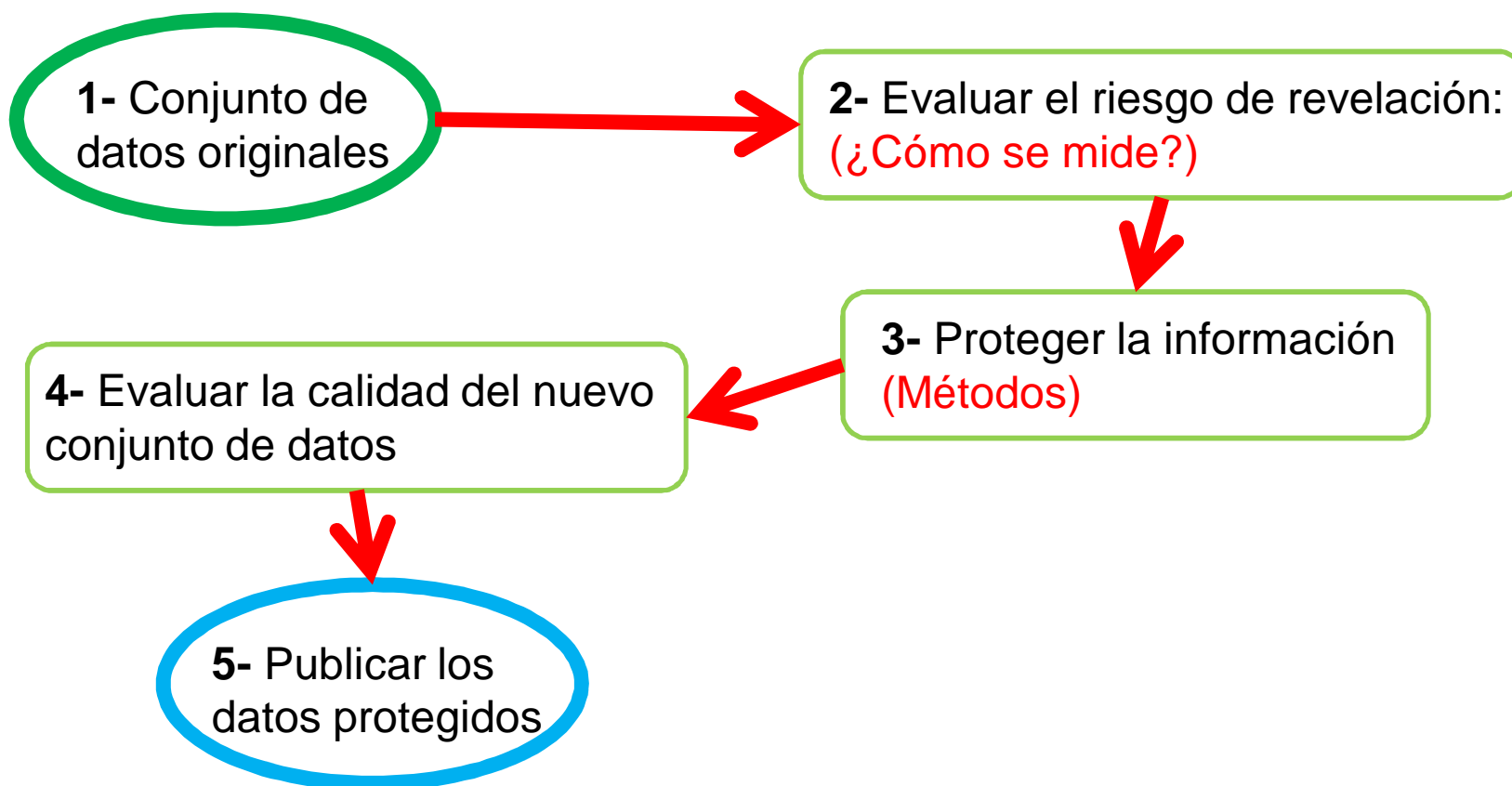
Magnitud: cada celda representa la suma de los valores de los informantes que contribuyen en esa celda

Frecuencia: cada celda representa el número de informantes que cumplen las características de tal celda

	Frecuencia	Magnitud	
	Nº Empresas	Volumen negocio	Gastos de personal
Hostelería	283.868	58.229.416	18.450.239
Información y comunicaciones	48.268	81.819.008	18.059.555

Protección de la información

Esquema general de trabajo



Protección de la información

Enmascarar los datos originales

Métodos de perturbación

Alterar los datos

Métodos de restricción

Ocultar los datos

Generar datos sintéticos

Crear datos *artificiales* que conserven ciertas propiedades estadísticas

Deseable: que se conserven propiedades estadísticas



Microdatos: evaluar

¿Cómo se reconoce a un individuo de entre los miles de registros que puede tener un fichero de microdatos?→ Informantes con características muy especiales (*combinaciones raras*).

A tener en cuenta: los ficheros muy detallados y de datos relativamente actuales facilitan la identificación



Microdatos: evaluar

La evaluación del riesgo se basa en buscar combinaciones **raras**, poco frecuentes en la población.

Se define una combinación de variables, **variables clave**, y se establece que si tal combinación de valores se da menos de x veces en la población, es una combinación insegura. (Ejemplo: edad < 40 estado civil =viudo/a tamaño del municipio<1000)



Microdatos: evaluar

Establecer tales combinaciones de variables: **expertos en la encuesta.**

No hay reglas preestablecidas. Por eso documentar las decisiones es útil porque pueden ayudar a la creación de buenas prácticas.



Microdatos: evaluar

Problemas:

Definición de *raro* / combinación de variables clave

Si estamos ante los microdatos de un censo, definir *raro* en el sentido de *poco frecuente*, es fácil. Pero

RARO en la muestra  *RARO* en la población



Microdatos: proteger

Data swapping y rank swapping

Intercambiar los valores de las variables confidenciales entre registros. No se cambian todos los registros, solo un porcentaje (pequeño) de ellos. Los registros deben ser *parecidos*.

Redondeo aleatorio:

Redondear el dato (cuantitativo) al múltiplo de una base escogida. No recomendable como único método de protección, siempre como refuerzo



Microdatos: proteger

Adición de ruido (noise addition): alterar los datos añadiendo un componente aleatorio

Microagregación: reemplazamiento de valores individuales con valores medios calculados a partir de agregaciones *pequeñas*.

Remuestreo: reemplazamiento de valores de una variable con medias obtenidas a partir de muestras de esa variable

PRAM (Post-Randomization Method): cambiar los valores de la variable usando una matriz de Markov (solo v. cualitativas)



Microdatos: proteger

Recodificación global

Data una variable, se combinan varias categorías para formar una nueva categoría menos específica.

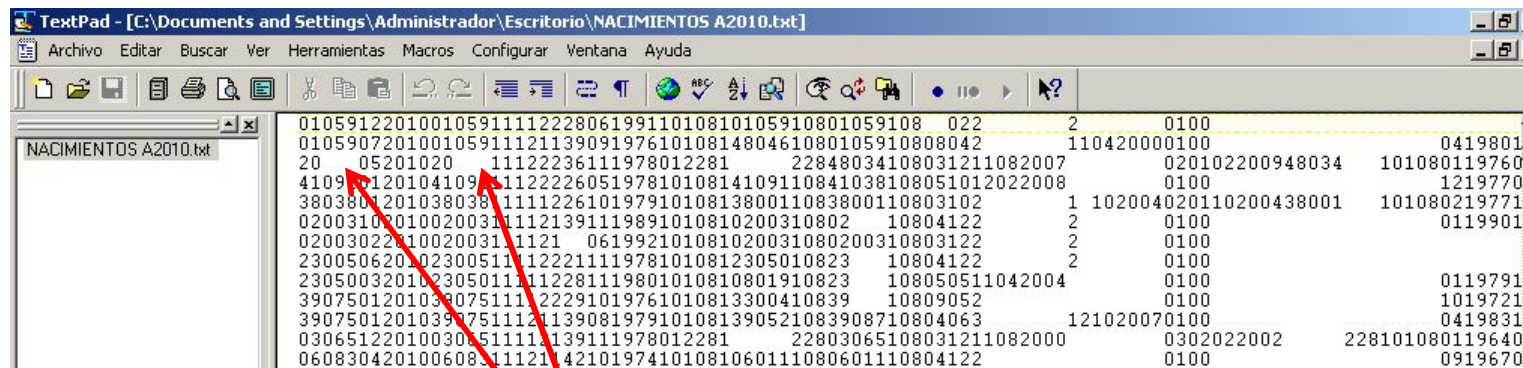
Es un método que se aplica a **todo** el archivo, no solo a la *parte no segura*.



Microdatos: proteger

Supresiones locales

Consiste en suprimir ciertos valores individuales de algunas variables



```
TextPad - [C:\Documents and Settings\Administrador\Escritorio\NACIMIENTOS A2010.txt]
Archivo  Editor  Buscar  Ver  Herramientas  Macros  Configurar  Ventana  Ayuda

0105912201001059111122280619911010810105910801059108  022  2  0100
010590720100105911121139091976101081480461080105910808042  110420000100  0419801
20  05201020  11122236111978012281  22848034108031211082007  020102200948034  101080119760
4109  120104109  11222260519781010814109110841038108051012022008  0100  1219770
380380  201038038  1111226101979101081380011083800110803102  1  102004020110200438001  101080219771
0200310  010020031  1121391119891010810200310802  10804122  2  0100  0119901
02003022  10020031  1121  061992101081020031080200310803122  2  0100
230050620  02300511  122211119781010812305010823  10804122  2  0100
2305003201  2305011  112281119801010810801910823  108050511042004  0100  0119791
3907501201039075111  22291019761010813300410839  10809052  0100  1019721
390750120103907511121139081979101081390521083908710804063  121020070100  0419831
030651220100306511111139111978012281  22803065108031211082000  0302022002  228101080119640
060830420100608  11121  42101974101081060111080601110804122  0100  0919670
```

Fichero de microdatos de nacimientos (MNP) con campos blanqueados



Microdatos: proteger

Importante: A veces la supresión da pistas sobre lo suprimido

Sexo	Profesión
Mujer	Matrona / matrón
-	Matrona / matrón
Mujer	Matrona / matrón
Mujer	Matrona / matrón



Tablas: evaluar

A tener en cuenta:

Tablas: conjunto de agregados creados a partir de los microdatos

Muchas veces contienen además subtotales y totales (sumas por filas o columnas y sumas de resultados marginales)



Tablas: evaluar

Ojo a...

La sensibilidad de las variables y la sensibilidad de algunos valores de las variables

Información muy fragmentada: tablas con muchos cruces, variables con muchas categorías



Tablas: evaluar

Reglas más frecuentes de evaluación del riesgo en datos tabulares:

1- Regla de la mínima frecuencia: se establece un parámetro n . Se considera que una celda es insegura si hay n o menos informantes que caigan en esa celda. Mínimo: $n = 2$ (celdas **seguras** si caen **3 o más informantes**).

Se usa cuando basta con prevenir que pueda deducirse el dato exacto

Variables cuantitativas y cualitativas



Tablas: evaluar

2- Regla de la dominancia: se establecen dos parámetros, n y k . Se considera que una celda es sensible si los n mayores contribuyentes aportan más del $k\%$ del valor de la celda.

Se usa cuando queremos dar más protección la celda: que el usuario no pueda deducir el valor exacto ocultado, ni tampoco hacer una buena estimación del mismo

Solo para variables cuantitativas



Tablas: evaluar

1- **Regla de la mínima frecuencia**: se escoge un parámetro n . Se considera que una celda es insegura si hay n o menos informantes que caigan en esa celda. Mínimo: $n = 2$ (celdas **seguras** si caen **3 o más informantes**).

2- **Regla de la dominancia**: se definen dos parámetros, n y k . Se considera que una celda es sensible si los n mayores contribuyentes aportan más del $k\%$ del valor de la celda.



Tablas: evaluar

A partir de los siguientes microdatos:

Empresa	Cifra de negocio	Rama actividad
Empresa 1	1200	A
Empresa 2	2000	A
Empresa 3	10	A
Empresa 4	20	A
Empresa 5	3600	B
Empresa 6	3500	B



Tablas: evaluar

Creamos la tabla “Cifra de negocio por rama de actividad”

Rama	Cifra de negocio
A	3230
B	7100

¿Es segura esta tabla?



Tablas: evaluar

Regla mínima frecuencia con $n=2$

Rama	Cifra de negocio
A	3930
B	7100

Caen 4 empresas, ¡OK!

Solo caen 2 informantes
(empresas 5 y 6)

Regla de la dominancia con $n=1$, $k=85\%$

Rama	Cifra de negocio
A	3230
B	7100

$61,92 < 85$

$50,7 < 85$



Tablas: evaluar

Regla dominancia con $n=2$, $k=85\%$

Rama	Cifra de negocio
A	3230
B	7100

$99,07 > 85$

$100 > 85$



Tablas: proteger

Al igual que con los microdatos, distinguimos entre métodos de perturbación y de restricción. Algunos de los más habituales son :

Redondeo, adición de ruido (de perturbación)

Supresión de celdas, recodificación (de restricción)

Las tablas tienen el añadido de que, al proteger, hay que tener en cuenta la **existencia de relaciones matemáticas entre los datos publicados**. Básicamente, que la suma de varios elementos de la tabla da otro elemento.



Tablas: proteger

Recodificación global: al igual que para los microdatos, consiste en combinar varias categorías para formar una más general.

Ejemplos: dar la edad por grupos quinquenales en lugar de año a año, la rama de actividad a dos dígitos en lugar de más, la nacionalidad por continente en lugar de por países, datos provinciales en lugar de municipales, combinar las categorías “divorciado” y “viudo” en “divorciado o viudo” etc.



Tablas: proteger

Supresión local: eliminar la información sensible y reemplazarla por *missing*. Normalmente se indica que es un *missing* por secreto estadístico.

Encuesta de financiación y gastos de la enseñanza privada. Curso 2009-2010

ENSEÑANZA UNIVERSITARIA. RESULTADOS CORRIENTES

Resultados corrientes (Valores absolutos) por Comunidades y Ciudades Autónomas y Nivel educativo

Unidades: miles de euros

	Estudios de Grado y de 1er y 2º ciclo	Estudios de Máster Oficial y Doctorado	Estudios propios No Oficiales
	Ingresos corrientes	Ingresos corrientes	Ingresos corrientes
Andalucía	29.512
Aragón
Asturias, Principado de
Baleares, Illes	5.355
Canarias
Cantabria

Notas:

1) '..' Dato protegido por secreto estadístico

.. El valor del dato es cero

Fuente: Instituto Nacional de Estadística



Tablas: proteger

Debido a la existencia de totales en las tablas, no basta con suprimir las celdas con información confidencial (supresión primaria). Hay que suprimir otras celdas para evitar que se pueda obtener la información suprimida restando, del total, la información disponible (supresión secundaria).

La supresión primaria y secundaria óptima, aquella que conduce a una mínima pérdida de información, requiere la resolución de un problema de programación lineal que no es trivial.



Tablas: proteger

Problema: si hay que suprimir muchas celdas no es el método más adecuado (tablas llenas de *missings*)

El método no protege contra la diferenciación. Si hay que proteger varias tablas que se elaboran a partir de la misma fuente, puede ser complicado gestionarlas todas para que se oculte el mismo dato en todas.

Ojo al proteger los **ceros confidenciales**

Ojo con los totales que revelen información de colectivos

Al publicar la tabla no se debe distinguir si la supresión es primaria o secundaria



Tablas: proteger

Empresa	Cifra de negocio	Rama actividad
Empresa 1	1200	A
Empresa 2	2000	A
Empresa 3	10	A
Empresa 4	20	A
Empresa 5	3600	B
Empresa 6	3500	B
Empresa 7	7800	C
Empresa 8	8200	C
Empresa 9	8000	C
Empresa 10	7900	C



Tablas: proteger

Rama	Cifra de negocio
A	3230
B	7100
C	31900
Total	42230

Evaluamos usando la regla de la mínima frecuencia, $n=2$

Rama	Cifra de negocio
A	3230
B	7100
C	31900
Total	42230



Tablas: proteger

Decidimos no publicar la celda sensible, y publicamos esto...

Rama	Cifra de negocio
A	3230
B	-
C	31900
Total	42230

... que es como si no hubiéramos hecho nada:

$$B = 42230 - 3230 - 31900$$

Debido a la presencia de totales es necesario ocultar algo más:
Supresión **primaria** y supresión **secundaria**



Tablas: proteger

Tenemos varias posibilidades

Rama	Cifra de negocio
A	-
B	-
C	31900
Total	42230

Rama	Cifra de negocio
A	3230
B	-
C	-
Total	42230

Rama	Cifra de negocio
A	3230
B	-
C	31900
Total	-

¿Cuál es la mejor?

Aquella en la que pierda menos información...

Si se puede, se recomienda no ocultar los totales



Tablas: proteger

Empresa	Cifra de negocio	Rama actividad	Región
Empresa 1	1200	A	Norte
Empresa 2	2000	A	Sur
Empresa 3	10	A	Este
Empresa 4	20	A	Oeste
Empresa 5	3600	B	Norte
Empresa 6	3500	B	Oeste
Empresa 7	7800	C	Sur
Empresa 8	8200	C	Sur
Empresa 9	8000	C	Este
Empresa 10	7900	C	Norte



Tablas: proteger

Cifra de negocio por rama de actividad y región

	Rama A	Rama B	Rama C	Total
Norte	1200	3600	7900	12700
Sur	2000	-	16000	18000
Este	10	-	8000	8010
Oeste	20	3500	-	3520
Total	3230	7100	31900	42230



Tablas: proteger

Todas las celdas son inseguras usando la regla de la mínima frecuencia $n=2$

	Rama A	Rama B	Rama C	Total
Norte	1200	3600	7900	12700
Sur	2000	-	16000	18000
Este	10	-	8000	8010
Oeste	20	3500	-	3520
Total	3230	7100	31900	42230



Tablas: proteger

Solución 1: supresión de celdas

	Rama A	Rama B	Rama C	Total
Norte	.	.	.	12700
Sur	.	.	.	18000
Este	.	.	.	8010
Oeste	.	.	.	3520
Total	3230	7100	31900	42230

Claro que para publicar esto, directamente no cruzamos las tablas



Tablas: proteger

Solución 2: construyo dos tablas sin cruzar

	Total
Norte	12700
Sur	18000
Este	8010
Oeste	3520
Total	42230

	Total
Rama A	3230
Rama B	7100
Rama C	31900
Total	42230

¡Ojo! Habría que volver a evaluar las celdas de estas nuevas tablas



Prácticas del INE

Dependen de cada operación estadística

Las prácticas comunes a todas ellas:

Identificadores directos eliminados lo antes posible

Se controla el nivel de desagregación geográfica (pocas operaciones con datos a nivel municipal), de ocupación, de actividad económica etc.

Prácticas del INE

En casi todas las ocasiones la protección no se reduce a la aplicación de un método, sino que se combinan varios: recodificación global y supresión de celdas, número de variables en cada tabla, control del número y detalle de las variables si la elaboración de tablas es “a la carta” (DWH Censos).

En microdatos se controla sobre todo el nivel de detalle geográfico, y dependiendo de la operación, otras variables como la ocupación, edad, nacionalidad, actividad económica, fechas, etc. Para su anonimización se usa sobre todo la recodificación global y la supresión de celdas.



Prácticas del INE

Además:

Para resolver dudas, elaborar recomendaciones sobre SE, gestionar las solicitudes de acceso a datos confidenciales por parte de terceros etc., el INE tiene un **Comité de Secreto Estadístico (CoSE)**.

Integrado por 9 personas

Todas las unidades del INE representadas



Acceso a microdatos confidenciales

Acceso a microdatos confidenciales con fines científicos



Acceso a microdatos confidenciales

Problema: a veces los investigadores necesitan datos (microdatos) más detallados de lo que publicamos para todo el mundo.

“Más detallados” significa que llegan al punto de ser confidenciales

Conflicto

Utilidad de la información – Protección de la información



Acceso a microdatos confidenciales

Política general: se debe facilitar al investigador el acceso a microdatos confidenciales con fines científicos.

Marco legal:

Ley de la Función Estadística Pública 1989 - Ponencia de Protección de Datos Estadísticos del Consejo Superior de Estadística (1995): la investigación con datos estadísticos puede entenderse como una fase más del proceso estadístico, y los investigadores considerarse personal estadístico (sujetos al deber de secreto estadístico por el artículo 17.3 de la LFEP).

Reglamento europeo 223/2009 y 557/2013



Acceso a microdatos confidenciales

Al final:

Puede permitirse el acceso a microdatos confidenciales bajo ciertas condiciones

- Solo se permitirá a investigadores que colaboren con instituciones de investigación.
- Solo se permitirá el acceso con fines científicos
- Los investigadores quedan sujetos a preservar la confidencialidad de los datos a los que tengan acceso
- Se da la información mínima necesaria para satisfacer la petición



Acceso a microdatos confidenciales

Proceso de solicitud:

- 1- El solicitante debe cumplimentar un modelo de solicitud explicando en qué consiste su proyecto, las variables que requiere, la justificación de la necesidad de acceder a datos confidenciales para elaborar el estudio y las medidas de seguridad que se aplicarán a la información.
- 2- Se revisa la solicitud (CoSE) y si procede se prepara la documentación pertinente para facilitar el acceso.
- 3- El acceso requiere la firma de un protocolo y de compromisos individuales e institucionales de confidencialidad.



Acceso a microdatos confidenciales

Nunca se permite el acceso a microdatos con datos de identificación directa.

Existen distintas modalidades de acceso en función del nivel de detalle solicitado:

- Envío de los microdatos al solicitante
- Acceso en centro seguro
- Ejecución de rutinas



Acceso a microdatos confidenciales

Caso especial: la investigación epidemiológica

Se podrá añadir información a los ficheros utilizados en investigaciones epidemiológicas que requieran conocer el estado vital y la causa de la muerte, o datos sobre la localización de la residencia, de los integrantes de la cohorte que se esté estudiando.



Acceso a microdatos confidenciales

Condiciones:

Solo se añadirá información de estado vital – causa muerte
– localización de residencia de personas fallecidas

El cruce de ficheros se hará en el INE, y lo hará personal
del INE

El fichero debe estar previamente inscrito en la AEPD

El cruce requerirá la firma de protocolo

Software específico

Software específico



Software específico

μ -argus y τ -argus son dos programas desarrollados en la oficina de estadística de los Países Bajos, dentro del programa CASC (Computational Aspects of Statistical Confidentiality), usados para ayudar a la anonimización, respectivamente, de micro y macrodatos.

Llevan implantados varios de los métodos expuestos aquí de evaluación de riesgo, protección y evaluación de pérdida de información.



Software específico

Además:

Algoritmo de supresión de celdas de la Agencia Tributaria

<http://administracionelectronica.gob.es/es/ctt/jcsp>

Soluciones en R

<http://dl.acm.org/citation.cfm?id=1556441>



Bibliografía

Bibliografía y documentos de interés



Bibliografía

Software μ -argus y τ -argus, proyectos CASC, CENEX, ESSnet, manuales, casos resueltos y ejemplos:

<http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cindex.htm>

Página de la UNECE con documentos sobre confidencialidad estadística:

<http://www.unece.org/stats/confidentiality.html>



Bibliografía

Legislación española:

Ley de la Función Estadística Pública (1989)

<http://www.boe.es/boe/dias/1989/05/11/pdfs/A14026-14035.pdf>

Ley Orgánica de Protección de Datos de carácter personal (1999)

<http://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf>



Bibliografía

Legislación europea:

Reglamento 223/2009 relativo a la Estadística Europea

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:es:PDF>

Reglamento 557/2013 sobre el acceso a datos confidenciales con fines científicos

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:ES:PDF>



Gracias por vuestra atención

alicia.fernandez.sanz@ine.es

