ESTADÍSTICA ESPAÑOLA
Vol. 38, Núm. 141, 1996, págs. 219 a 235

Estimación de los DNI duplicados en España

por
JUSTINO GARCÍA DEL VELLO
Doctor Ingeniero de Caminos, Canales y Puertos
y Administrador Civil del Estado
(justino.garcia@dgopti.map.es)

RESUMEN

El número del Documento Nacional de Identidad (DNI) no es una clave perfecta y muchos españoles lo tienen duplicado. Los errores de transcripción son la causa más frecuente de estas duplicaciones y son difíciles de detectar, puesto que de ellos no queda constancia en los registros oficiales. Se presenta una estimación de los DNI duplicados en manos de españoles vivos, y se ofrece un modelo estadístico que fácilmente puede usarse en otras estimaciones. Como aportación más trascendente, se demuestra que las estimaciones oficiosas, que tradicionalmente se han elaborado con información de grandes bases de datos no exentas de ruido, sistemáticamente están sesgadas y agrandan la cifra real de duplicados. Por último, se muestra un procedimiento para estimar el ruido de una base de datos a partir de las duplicaciones internamente observadas de un identificador no repetible.

Palabras clave: Documento Nacional de Identidad, DNI, duplicación, error de transcripción, depuración de bases de datos.

Clasificación AMS: 62P99.

1. PANORÁMICA

1.1. Introducción al problema

Harto conocido por numerosos colectivos profesionales es el hecho de que el número del Documento Nacional de Identidad (DNI) no es una clave perfecta: muchos españoles comparten su número del DNI con otros españoles. La duplicación del número del DNI tiene especial relevancia porque provoca serios problemas en la vida civil (bancos, grandes empresas de servicios, recaudación de impuestos, etc). Estos problemas, sin embargo, no son originados por las tarjetas(1) del DNI que, aun portando un número erróneo, no causen duplicaciones.

No faltan especulaciones acerca de la cantidad de DNI duplicados. Las más optimistas la cifran en unos pocos millares y las más pesimistas la colocan rondando el millón. Sin embargo sorprende la ausencia de publicaciones de estudios bien documentados que, al menos, acoten su orden de magnitud. Ésta es la tarea que nos vamos a plantear.

Se suele creer que la cantidad de DNI duplicados ha de ser sobradamente conocida por la Dirección General de la Policía, puesto que conserva los registros de los DNI emitidos, pero ello no es así por varios motivos. El más importante de todos es que hay muchas tarjetas del DNI cuyos números no coinciden con los que aparecen en sus correspondientes registros oficiales. Además, parte de los registros oficiales de antaño se deterioraron con el paso de los años(2). Así pues, los DNI duplicados no son detectables simplemente estudiando los registros oficiales.

Conviene en este momento hacer una descripción somera de la "confección y expedición teórica" del DNI. Existen numerosos centros expedidores repartidos por la geografía nacional. Centralizadamente se otorgan a cada provincia grandes series de números consecutivos para ser asignados a los DNI. Cuando un ciudadano solicita su primer DNI, se crea un registro oficial nuevo de cierta complejidad, cuya descripción omitimos por innecesaria, y se le asigna el número secuencial que corresponda. A continuación se confecciona la que, impropiamente, venimos llamando tarjeta del DNI que se entrega al solicitante. Cuando el ciudadano solicita

---

(1) Vamos a usar la palabra *tarjeta*, pese a que en este contexto es inusual y algo redundante con *documento*, para evitar la ambigüedad que tiene la expresión *número del DNI*, pues hemos de distinguir entre el número que figura en los registros oficiales y el número que porta una *tarjeta* del DNI. Desgraciadamente no siempre coinciden.

(2) Desde hace unos años, el DNI se gestiona e imprime con la ayuda de sistemas de información automatizados que, prácticamente, eliminan estas fuentes de error.

la renovación, se comprueba la veracidad de su tarjeta vieja contrastándola con su registro oficial, y se confecciona una nueva tarjeta que la sustituirá.

¿Cuáles son los problemas de este procedimiento? Prácticamente ninguno desde que, ya en la década de los años noventa, se utilizan sistemas de información automatizados para gestionar y confeccionar los registros oficiales y las tarjetas, y se actualiza una nueva base de datos a nivel nacional que no admite duplicados y que sólo contiene los DNI nuevos o renovados en estos últimos años. Pero anteriormente, sin la ayuda de la informática, las cosas eran muy diferentes. Registros oficiales y tarjetas se cumplimentaban con máquina de escribir, y los llamados *errores de transcripción* (entiéndase de mecanografía y similares) eran inevitables. Por otro lado, no siempre era posible realizar la verificación de las tarjetas en las renovaciones (piénsese en la movilidad ciudadana, en una dramática escasez de medios y en una fortísima presión social para que los trámites se agilizaran), lo que ocasionaba una acumulación de errores en lugar de su corrección. Por último, hubo alguna *duplicación de series* al asignar ciertos centros expedidores los mismos números por equivocación.

En la práctica, la situación actual es relativamente simple de describir. Los errores detectables con los registros oficiales están bajo control y son irrelevantes comparados con los errores de transcripción que llevan las tarjetas del DNI de muchos españoles, que sólo podrán ser detectados con su renovación. Estos errores de transcripción son los responsables de la inmensa mayoría de los DNI con número duplicado actualmente existentes.

Errores como la repetición de series crean duplicados solamente durante la emisión de nuevos DNI (no así en las renovaciones), concentrándolos en períodos y puntos geográficos de emisión muy concretos, lo que hacen difícil su estimación estadística. Por el contrario, los errores de transcripción pueden surgir tanto en la emisión como en las renovaciones del DNI, y su aparición está mucho más uniformemente distribuida en el espacio y en el tiempo. Por ello su estimación estadística es, relativamente, más factible.

Este estudio se va a centrar fundamentalmente en la estimación de los duplicados causados por errores de transcripción. Sin embargo, siendo comparativamente irrelevante la cantidad de duplicados actuales originados por otras causas, podemos afirmar que la estimación que hagamos será representativa del número total de duplicados existentes.

1.2. Métodos alternativos para abordar el problema

Dado que los DNI duplicados no pueden ser "contados" en ningún libro oficial, no queda otro camino que estimarlos estadísticamente. Realizar un muestreo convencional "ad hoc" para hacer la estimación estadística sería poco menos que disparatado. Por un lado, el tamaño de la muestra habría de ser superior a 250.000 individuos (consecuencia, como se verá, de una conocida recomendación estadística). Por otro lado, a no ser que se tomaran medidas de seguridad excepcionales, el propio ruido del muestreo imposibilitaría una estimación fiable (tales son las tremendas dificultades que entraña jugar con una pregunta que tiene cien millones de respuestas posibles). Bien es cierto que habría otros caminos indirectos de muestreo más plausibles pero, en cualquier caso, habría de realizarse una importante inversión económica que no se justificaría en los tiempos que corren.

Rechazada la alternativa de un muestreo "ad hoc", el siguiente paso es indagar si alguna base de datos ya existente pudiera proporcionar la información necesaria para realizar la estimación. Esto no es nada fácil porque la base ha de cumplir severas condiciones. En primer lugar, ha de cumplir el requisito ya mencionado de contener datos de un colectivo no inferior a 250.000 personas. En segundo lugar, la base no debe contener ruido (esto es, equivocaciones fortuitas o voluntarias en los procesos de captura y mantenimiento de la información). Un pequeño nivel de ruido sería admisible si son sistemáticamente investigadas y documentadas en profundidad todas las duplicaciones que aparezcan en la base(3). Por último, el colectivo de personas de la base ha de ser representativo de la población española que posee DNI.

En rigor, ninguna base de datos existente satisface plenamente las tres condiciones exigidas. No obstante, algunas permiten hacer una aproximación al problema. Para nuestro estudio hemos elegido la base de datos del personal en activo

---

(3) Merece un comentario adicional el cuidado especial que debe tenerse con el ruido de la base de datos. El problema del ruido es muy grave porque **sesga** las estimaciones, **aumentando** sistemáticamente la cantidad aparente de DNI duplicados. Se demuestra en el estudio que los errores de la propia base en los números del DNI pueden crear más de doscientos falsos duplicados por cada pareja de duplicados auténticos que enmascaren. Además, otros tipos de ruido también sesgan los resultados en el mismo sentido. Por ejemplo: no dar de baja a los fallecidos; prácticas viciadas, como la asignación a los menores de edad del número del DNI de su padre; considerar dos registros de una misma persona como pertenecientes a dos personas distintas con el DNI duplicado cuando hay un error en algún otro dato identificador, como apellidos, nombre o fecha de nacimiento; por último, no deben ser olvidados los errores deliberados que originan los intentos de fraude. Es por ello que las estimaciones que a veces se han hecho a partir de los duplicados aparentes que hay en algunas grandes bases de datos están siempre sesgadas y sistemáticamente agrandan la cifra real de duplicados.

inscrito en el Registro Central de Personal (RCP) del Ministerio de Administraciones Públicas.

1.3. Conclusiones obtenidas

1. La emisión de nuevos DNI con su número correcto también aumenta la cantidad de DNI duplicados, por extraño que parezca.

2. La inmensa mayoría de los DNI repetidos está formada por duplicados, siendo muy raros los triplicados. A su vez, la inmensa mayoría de estas duplicaciones está formada por una tarjeta con el número del DNI correcto y otra que lo tiene erróneo. Son muy poco frecuentes las duplicaciones formadas por dos tarjetas que tienen ambas el número del DNI equivocado.

3. Existe una apreciable cantidad de tarjetas cuyo número, aun sin estar duplicado, también tiene errores de transcripción. Son la principal causa de que la emisión de nuevos DNI correctos también aumente la cantidad de DNI duplicados.

4. Como consecuencia de lo anterior se desprende que para solucionar de una vez por todas el problema de las duplicaciones no basta con corregir éstas: es condición necesaria retirar de la circulación todas las tarjetas con número de DNI erróneo. Una alternativa cuando el error no ha causado todavía duplicación consiste en modificar los registros oficiales, evitando así las molestias del cambio al ciudadano.

5. Carecemos de datos explícitos para estimar estadísticamente el número de DNI duplicados por causas diferentes a los errores de transcripción (como repetición de series y otras). No obstante, por diversas conjeturas e informaciones parciales sobre el tema, podemos afirmar que la cantidad de estos duplicados es irrelevante comparada con la causada por errores de transcripción.

6. Los errores de transcripción no siguen una distribución uniforme, esto es, cuando se producen tienen tendencia a cambiar el número correcto en otro número "parecido" y no en cualquier número arbitrario. Así, la mayoría de los errores de transcripción afectan a un solo dígito, siendo el más frecuente el primero de la derecha. Este comportamiento tiene gran importancia en la distribución de los duplicados, pues teniendo cada provincia otorgadas grandes series de números consecutivos para asignar a los DNI, es muy probable que el error de transcripción produzca un número (erróneo) que también pertenezca a la misma provincia.

7. En base a las parejas de duplicados observadas en el Registro Central de Personal de los DNI del personal en activo de la Administración inscrito en el mismo, y asumiendo la hipótesis de que los errores de transcripción siguieran una distribución uniforme, podría fijarse el valor esperado de la cantidad de DNI dupli-

cados por error de transcripción dentro del colectivo de españoles vivos, en 1996, en unas 105.000 unidades (lo que corresponde a un 0,32 % de los DNI en manos de dicho colectivo). Ahora bien, por los motivos expresados en el apartado anterior y por no estar el colectivo en estudio repartido equilibradamente por provincias, podemos afirmar que dicha cifra debe revisarse notablemente a la baja. No disponemos de información suficiente para poder cuantificar en qué magnitud, pero bien podemos afirmar que la cantidad real buscada de DNI duplicados viene expresada con **cinco dígitos** solamente. Recordemos, a su vez, que la mitad de los DNI duplicados son correctos.

8. La cifra de DNI duplicados en España que presentamos, lejos de ser una mala noticia, es una noticia excelente, pues las estimaciones oficiosas que se han venido haciendo tradicionalmente son muy superiores. La causa ha sido que se han realizado con información de grandes bases de datos no exentas de algún ruido. Y, según se demuestra en el estudio, el ruido crea muchos más falsos duplicados en la base que los verdaderos que oculta. De ahí que estas estimaciones estén siempre **sesgadas** y **agranden** sistemáticamente la cantidad buscada. La única forma de evitar este sesgo es realizar previamente una documentación rigurosa y exhaustiva de todos los duplicados observados, contrastándolos con los documentos originales. Posiblemente sea ésta la conclusión más trascendente de nuestro estudio.

9. La corrección de los DNI está mejorando notablemente gracias a la implantación de sistemas de información y de gestión automatizados a comienzos de la década de los noventa, que reducen drásticamente los frecuentes e inevitables errores del proceso manual de antaño. A comienzos de la próxima década (que será también comienzos de siglo y de milenio), debido a las renovaciones preceptivas, serán meramente testimoniales los errores y duplicaciones en los números de los DNI en circulación. Sólo un pequeño colectivo seguirá, mientras viva, manteniendo abierto el problema: los actualmente mayores de 70 años, que no están obligados a renovar su DNI.

10. Tal es la importancia del sesgo que venimos comentando, que uno de los modelos estadísticos desarrollados en este estudio permite obtener, en grandes bases de datos, una primera aproximación rápida y barata del porcentaje de los DNI almacenados con errores en la propia base. Para ello basta aplicar una fórmula a la cantidad de DNI duplicados que se observen en la misma base. Este procedimiento es fácilmente generalizable a cualquier base de datos que contenga algún atributo concebido para servir de identificador único de personas, objetos, situaciones, permisos y demás. Por ejemplo: matrículas de coche, números de afiliación a la Seguridad Social, números de identificación en el Censo Electoral, NIF, CIF, claves de licencias diversas, etc.

2. MODELO ESTADÍSTICO UTILIZADO

Como se indicó previamente, se ha elegido la base de datos de personal en activo inscrito en el Registro Central de Personal (RCP) del Ministerio de Administraciones Públicas(4).

Esta base tiene un tamaño suficiente(5) para realizar la estimación estadística que buscamos; posee un nivel de ruido muy pequeño (menor del uno por mil) en los datos que nos interesan a causa de las continuas depuraciones y muestreos que se vienen realizando para garantizar la calidad de información; y, más importante aún, todas las duplicaciones de DNI observadas desde 1991 son sistemáticamente investigadas y documentadas(6), por lo que se puede totalmente rechazar la hipótesis de un engrosamiento de los DNI duplicados por ruido en la información.

Con todo, también el RCP tiene sus inconvenientes. Sin duda, el inconveniente *a priori* más serio es que el colectivo escogido no es una muestra representativa de la población española, hablando en un sentido amplio. Sin embargo, como veremos, se juzga que es suficientemente representativa en los aspectos que afectan a nuestro problema como para permitir acotar el problema.

El estudio estadístico realizado infiere los DNI duplicados por error de transcripción(7) dentro del colectivo de españoles vivos, a partir de las parejas de duplicados internamente observados dentro del personal en activo contenido a comienzos de 1996 en el RCP. Tres hipótesis subyacen en nuestro proceder:

---

(4) Hemos optado por la base de datos del RCP porque hemos conseguido las autorizaciones oportunas para publicar los resultados. Como algunas otras bases de datos también podrían servir para investigar el tema que nos ocupa, publicamos con cierto detalle los cálculos realizados. Así, bastará sustituir las variables por sus nuevos valores para obtener nuevas estimaciones.

(5) Como fácilmente podrá comprobar el experto, la universalmente aceptada recomendación del tamaño mínimo de una muestra para estimar una proporción [$n_p \cdot p_p \geq 5 \leq n_p \cdot q_p$] conduce, en nuestro caso, a precisar un colectivo de, al menos, 250.000 personas.

(6) Debe señalarse que, desde 1991, toda duplicación de números de los DNI del personal inscrito en el RCP es contrastada documentalmente y, una vez confirmada, es comunicada a la Dirección General de la Policía para su oportuna corrección. Estas duplicaciones nos proporcionan una excelente oportunidad para realizar una estimación de los DNI duplicados a nivel nacional.

(7) La cantidad de duplicados debidos a otras causas (como repetición de series) es irrelevante comparada con la debida a errores de transcripción. De hecho, todos los duplicados observados internamente en el RCP corresponden a esta categoría. Además, de las rectificaciones oficiales de DNI recibidas en el RCP, más del 97 por ciento tiene por causa errores de transcripción obvios.

Primera hipótesis: los errores de transcripción que se han producido en los DNI son independientes entre sí, tanto geográfica como temporalmente.

Segunda hipótesis: el mecanismo que genera el fenómeno de convertirse en funcionario o trabajador laboral de la Administración se va a suponer que nos proporciona una muestra aleatoria simple (esta hipótesis es válida sólo en la medida en que lo sea la anterior).

Tercera hipótesis: los errores de transcripción, una vez producidos, convierten aleatoriamente un número de DNI en otro siguiendo una distribución uniforme. Esta última hipótesis no es cierta, pues sabemos que la mayoría de los errores afectan a un solo dígito, siendo el más frecuente el primero de la derecha. Teniendo cada provincia otorgadas grandes series de números consecutivos para asignar a los DNI, es muy probable que el error de transcripción produzca un número (erróneo) que también pertenezca a la misma provincia. Comoquiera que el colectivo en estudio no está repartido equilibradamente por provincias (y no disponemos de información suficiente para proceder de otra manera), la estimación que se va a obtener de duplicados será superior a la real(8). En cualquier caso, esta cota superior obtenida (o cifra de duplicados claramente superior a la real) va a ser de gran utilidad porque, a su vez, es muy inferior a las estimaciones oficiosas que se han venido realizando utilizando grandes bases de datos no exentas de cierto ruido.

2.1. Explicación para los no familiarizados con la Estadística(9)

Llamemos N (tamaño de la población) al número de españoles vivos con DNI.

Llamemos n (tamaño de la muestra) al número de personas que forman nuestra muestra. Supondremos que n es muy grande y que la muestra es representativa de la población.

Llamemos M al número de españoles cuyo DNI tiene el número duplicado a causa de errores de transcripción. Supondremos que no existen números de DNI triplicados porque su cantidad es muy pequeña comparada con la de duplicados, y porque esta suposición simplifica notablemente el problema.

---

(8) Debido a que la cantidad de DNI duplicados internamente observados crece más rápidamente que el tamaño de la muestra.

(9) Se incluye este insólito apartado para advertir a las personas no familiarizadas con la Estadística que no es correcto aplicar la *regla de tres* ni siquiera para obtener una estimación aproximada. Además, se pretende dar un método alternativo que, sencillo y robusto, permita confirmar la corrección de la solución al experto que no disponga del tiempo y energías precisas para calentarse la cabeza con la solución formal.

Vamos a descomponer nuestro problema en cuatro cuestiones más sencillas.

• PRIMERA CUESTIÓN: ¿Cuántos DNI cuyo número esté repetido con el de algún otro español habrá en la muestra? Evidentemente, se espera que haya M.n/N. Sin embargo, esto no nos resuelve el problema. En efecto, no podemos identificarlos simplemente estudiando la muestra porque los otros DNI que duplican su número habrán quedado, en la mayoría de los casos, fuera de la muestra.

• SEGUNDA CUESTIÓN: ¿Qué probabilidad tiene un DNI incluido en la muestra y cuyo número esté repetido con el de algún otro español de que su pareja que lo duplica también se encuentre en la muestra? Obviamente, dicha probabilidad (casos favorables partido por casos posibles) es (n-1) / (N-1)

• TERCERA CUESTIÓN: ¿Cuál es el número Y de DNI duplicados que se espera que estén en la muestra junto a su pareja que lo duplica? Bastará multiplicar el resultado de la primera cuestión por el resultado de la segunda. Por lo tanto, Y = M . n . (n-1) / (N . (N-1))

• CUARTA CUESTIÓN: ¿Cuál será el valor esperado de la cantidad M de españoles cuyo DNI tiene el número duplicado, si en la muestra aparece un número Y de DNI que están duplicados con otros DNI de la misma muestra? Despejando M en la solución de la tercera cuestión: M = Y . N . (N-1) / (n . (n-1))

El colectivo elegido en nuestro estudio es el personal en activo a comienzos de 1996 inscrito en el RCP. Dando valores a las variables:

N = 33.251.098

n = 522.142

Y = 26

Obtenemos la solución: M = 105.440

Resumiendo y redondeando, se estima que hay unos 105.000 españoles cuyo DNI tiene el número duplicado por error de transcripción. (Recordemos que por no cumplirse la tercera hipótesis mencionada, la cifra real ha de ser inferior a la obtenida).

2.2. Cálculos estadísticos formales

Sea N (tamaño de la población) el número de españoles vivos con DNI; sea n el tamaño de la muestra elegida, que supondremos representativa de la población; y sea M el número de españoles vivos cuyo DNI tiene el número duplicado a causa de errores de transcripción. Supondremos que no existen números de DNI triplicados porque su cantidad es muy pequeña comparada con la de duplicados, y porque esta suposición simplifica notablemente el problema.

Fijemos nuestra atención en el conjunto de todas las parejas posibles que pueden formarse con los DNI que poseen los españoles vivos. Su cardinal N_p (combinaciones de N elementos tomados dos a dos) valdrá: $N_p = N \cdot (N-1) / 2$.

Del conjunto anteriormente definido, el número M_p de parejas cuyos dos elementos (DNI) tienen el mismo número será: $M_p = M / 2$.

Análogamente, con una muestra de n elementos (DNI) elegidos al azar, el número n_p de parejas de elementos que se pueden obtener es: $n_p = n \cdot (n-1) / 2$.

Con este pequeño artificio hemos reconducido nuestro problema a otro cuya solución es inmediata. En efecto, tenemos una población de N_p elementos (parejas), M_p de los cuales poseen una característica que los diferencia de los demás (los dos números son iguales). De dicha población extraemos una muestra de n_p elementos, entre los que aparecerán Y_p elementos (parejas) con los dos números iguales. Se quiere estimar M_p a partir de la muestra. Ahora sí que estamos ante un tema clásico de la Estadística: estimación de una proporción (o, si se prefiere, de una binomial).

La solución:

Sea $\hat{\pi}_p = Y_p / n_p$

y $\pi_p = M_p / N_p$

$$\pi_p \simeq \hat{\pi}_p \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}_p \cdot (1 - \hat{\pi}_p)}{n_p}}$$

El colectivo estudiado es el personal de la Administración en activo a comienzos de 1996 inscrito en el RCP(10). Dentro de este colectivo, que está formado por 522.142 personas(11), hay 26 que tienen el número del DNI duplicado dos a dos (es

(10) Obviamente, el colectivo elegido no es representativo en un sentido amplio de la población española en posesión del DNI. Sin embargo, quienes han hecho los DNI, y con ellos los errores de transcripción inevitables que conllevan, no son los portadores finales de los mismos, sino un pequeño colectivo de funcionarios encargado de esta responsabilidad: colectivo con formación homogénea, movilidad territorial y renovación natural con el paso del tiempo. Véanse los comentarios a las tres hipótesis realizadas.

(11) Para las personas no familiarizadas con este contexto, aclararemos que este colectivo está formado no sólo por oficinistas, sino también por carteros, maestros, profesores, catedráticos, personal de instituciones penitenciarias, etc.

decir, hay 13 parejas de DNI con el número repetido). Se supone que son 33.251.098 los españoles vivos que poseen DNI(12).

Dando valores a las variables:

N = 33.251.098

n = 522.142

Y = 26

Obtenemos:

N_p = 5,52818 E+14

n_p = 1,36316 E+11

Y_p = 13

$Z_{\alpha/2}$ = 1,96 para un 95 % de confianza, esto es, para α = 0,05(13)

$\hat{\pi}_p = Y_p / n_p$ = 13 / 1,36316 E+11 = 9,53667 E-11

π_p = 9,53667 E-11 ± 5,18419 E-11

Y ahora podemos obtener los valores que realmente nos interesan:

$\pi = M / N = 2 . M_p / N = 2 . \pi_p . N_{pp} / N = 2 . \pi_p . N . (N-1) / (2.N) =$

$\pi = \pi_p . (N-1)$ = 3,17105 E-3 ± 1,72380 E-3 = 0,317105 ± 0,172380 %

$M = \pi . N$ = 105.441 ± 57.318

Resumiendo y redondeando, este estudio estima que a comienzos de 1996, dentro del colectivo de españoles vivos, el porcentaje de tarjetas del DNI con el número duplicado a causa de errores de transcripción era de 0,317 ± 0,172 % para un nivel de confianza del 95 %. En otras palabras, el número de tarjetas del DNI

---

(12) Se estima como número de españoles en posesión del DNI el número de españoles mayores de 14 años, cifra, a su vez, obtenida mediante una ligera extrapolación del Anuario Estadístico 1995 del INE. Dos pequeños colectivos desafían esta suposición en sentidos opuestos: los españoles mayores de 14 años que no poseen todavía el DNI, y los menores de 14 años que ya lo poseen junto a los españoles que poseen más de un DNI.

(13) Recordamos a quienes hubieran preferido escoger $t_{\alpha/2}$ en lugar de $z_{\alpha/2}$ que, en este contexto, ambos tienen el mismo valor dado el elevado número de grados de libertad.

con el número duplicado(14) a causa de errores de transcripción dentro del colectivo de españoles vivos se estima en 105.000 ± 57.000 para un nivel de confianza del 95 %.

Por no cumplirse la tercera hipótesis previamente mencionada, la cifra real de duplicados ha de revisarse notablemente a la baja. No disponemos de información suficiente para cuantificar su magnitud, pero al menos podemos afirmar que su valor esperado se escribe con **cinco dígitos** solamente. Recordemos, a su vez, que la mitad de los DNI duplicados son correctos.

3. SESGO CREADO POR EL RUIDO

La experiencia demuestra, y el cálculo justifica, que en grandes bases de datos con ruido siempre existe una cantidad de DNI duplicados aparentes internos muy superior a la que les correspondería si no tuviesen tal ruido. Por consiguiente, las estimaciones de la cantidad de DNI duplicados en España que se hagan a partir de las duplicaciones aparentes observadas **estarán siempre sesgadas agrandando la cifra real de duplicados**. Este sesgo sólo puede evitarse realizando una documentación rigurosa y exhaustiva de todos los duplicados observados, contrastándolos con los documentos originales. Veamos en detalle la justificación de estas afirmaciones.

---

(14) De los DNI duplicados la experiencia confirma que, en general, de cada pareja uno es correcto y otro erróneo (o, lo que es lo mismo, es muy rara la duplicación entre DNI erróneos). Esto es debido a que la inmensa mayoría de los DNI es correcta: por consiguiente un DNI con número equivocado tiene mucha más probabilidad de chocar con un DNI correcto que con otro incorrecto. De lo dicho, se desprende del estudio que sería 52.700 la cantidad esperada de DNI correctos que tienen su número duplicado, y también 52.700 la cantidad esperada de DNI con error de transcripción que tienen su número duplicado. En cualquier caso, recordamos que estas cifras deben revisarse a la baja por los motivos previamente señalados, y constituyen simplemente sólidas cotas superiores.

Dado que la que la mitad de los DNI duplicados tienen el número correcto, se concluye que un colectivo, elegido al azar, cuyos DNI estuvieran exentos de errores ¡todavía tendría un porcentaje esperado de DNI duplicados (con otros del país) que sería la mitad del nacional!

También, obviamente, hay un conjunto de DNI erróneos que, sin estar todavía duplicados, están a la "expectativa" de duplicar nuevos DNI. Por eso la emisión de DNI perfectamente correctos ¡también aumenta la cantidad de DNI duplicados! Sólo la eliminación de los DNI actualmente incorrectos permitirá detener este crecimiento sistemático.

Si se cumpliera la tercera hipótesis, para estimar la cantidad total de DNI con errores de transcripción bastaría multiplicar la cantidad obtenida de DNI duplicados erróneos, 52.700, por $10^8/N$, siendo N la cantidad de españoles vivos en posesión del DNI. Por no cumplirse dicha hipótesis, esta estimación es groseramente superior a la realidad.

3.1. Ruido en la captación del DNI

Consideremos una base de datos que contiene ruido en los números del DNI, consecuencia de errores (accidentales o voluntarios) producidos durante la recogida y almacenamiento de la información. Supondremos que los errores siguen una distribución uniforme y que dicha base contiene un único registro con el DNI de cada persona de cierto colectivo.

Como el número del DNI tiene ocho dígitos, pueden existir hasta cien millones de números diferentes. Sea N la cantidad de españoles vivos en posesión del DNI y sea $t = N/10^8$. Sea π la proporción de tarjetas del DNI en manos de españoles vivos cuyo número está duplicado. Sea n la cantidad de DNI que hay recogidos en la base de datos (supondremos que n es grande); sea R la cantidad de dichos números que no coinciden con el que figura en la tarjeta del DNI de su titular; y sea $r = R / n$

Si en la base de datos no hubiera ruido, los DNI duplicados observados serían todos auténticos. La cantidad esperada es:

$$E = \pi \cdot n \cdot (n-1) / (N-1)$$

Sin embargo, de ordinario habrá en la base de datos cierto ruido propio, que mediremos por r (supondremos que r es pequeña). Este ruido r afecta al número aparente de DNI duplicados internos de la base en dos sentidos opuestos. De un lado disminuirá el número de duplicados auténticos observados, a causa de la alteración de algunos DNI, y esa pérdida P valdrá:

$$P = 2 \cdot r \cdot E$$

Por otro lado, el ruido creará una cantidad A de aparentes duplicados:

$$A = 2 \cdot r \cdot n \cdot (n-1) / 10^8$$

Así, el número esperado D de duplicaciones internas observadas en la base será:

$$D = E + A - P = (n \cdot (n-1) / (N-1)) \cdot (\pi + 2 \cdot r \cdot t - 2 \cdot r \cdot \pi)$$

Para nuestros propósitos, vamos a presentar D de otra manera más "didáctica".

Sea $k = 2 \cdot r \cdot n \cdot (n-1) / 10^8$

$$D = E + A - P = E + (2 \cdot r \cdot n \cdot (n-1) / 10^8) \cdot (1 - \pi / t) \quad [1]$$

$$D = E + k \cdot (1 - \pi / t)$$

Siendo $A = k$

y $P = k \cdot \pi / t$

El factor común k depende del tamaño y ruido de la base de datos que estemos considerando, pero el término discriminador (1 - π / t) no depende de ella, sino de la coyuntura nacional del momento (esto es, del número de españoles vivos con DNI, y de los duplicados que existen en los DNI de dicho colectivo).

Finalmente, obtengamos la siguiente razón:

$$A / P = 1 / (\pi / t) = t / \pi \approx 0{,}3325 / 0{,}00317 \approx 105$$

Esto quiere decir que el ruido en los DNI de una base de datos crea unos 210 aparentes duplicados (que no lo son realmente) por cada pareja de duplicados auténticos que enmascara. Luego queda demostrado que el ruido en los números de los DNI **sesga** las estimaciones, **aumentando** sistemáticamente la cantidad aparente de DNI duplicados.

3.2. Ruido en otros atributos identificadores

Cuando la base de datos considerada puede contener varios registros con el DNI de cada persona de cierto colectivo, pueden aparecer otros ruidos que también sesgan la estimación de duplicados en el mismo sentido.

Así, registros que provienen de una misma persona pueden ser atribuidos a supuestas personas distintas cuando alguno de ellos contiene errores en otros atributos identificadores como nombre, apellidos o fecha de nacimiento(15). En este caso, cada error supone la aparición de una nueva pareja de DNI supuestamente duplicados en la base.

3.3. Ruido en los DNI de los menores de edad

Otra causa de sesgo en las duplicaciones de los DNI son ciertas prácticas viciadas (si bien desde otros puntos de vista a veces puedan ser consideradas como las actuaciones más razonables en ciertos contextos). Un ejemplo clásico es la asignación a los menores de edad del número del DNI (y aun del NIF) de su padre cuando están envueltos en alguna actividad económica. Incluso agencias públicas se ven obligadas, como mal menor, a documentar así sus transacciones. Cada actuación

---

(15) A veces se utilizan algoritmos "inteligentes" para discriminar si nombres y apellidos no literalmente iguales corresponden o no a la misma persona. No obstante, el procedimiento nunca es perfecto y sistemáticamente agranda indebidamente la estimación de DNI duplicados. Podría decirse que se crea un ruido aparente en un atributo a causa del ruido de otro. Este tipo de error crea un sesgo muy serio, pues cada registro con DNI correcto que no sea atribuido a su dueño origina una duplicación aparente si existe también un registro sin errores.

en este sentido también implica una nueva pareja de DNI supuestamente duplicados en la base.

3.4. No dar de baja a los fallecidos

Algunas grandes bases de datos contienen muchos más DNI que españoles vivos hay con DNI. Esta diferencia proviene generalmente de los ruidos identificados en los apartados 3.2 y 3.3, y de la práctica viciada de no dar de baja a los fallecidos (no siempre se dispone de información para hacerlo). Esta práctica trae consigo el aumento de la población considerada, y con ello un incremento notable de los DNI duplicados. (En una primera aproximación, el número de duplicados esperados entre los DNI de un colectivo depende del cuadrado del tamaño de dicho colectivo).

3.5. Intentos de fraude

Por último, no deben ser olvidados los errores deliberados que originan los intentos de fraude. Con frecuencia ello trae que una misma persona aparezca con más de un DNI en algunas bases de datos. Obviamente, los DNI inventados también aumentan la probabilidad de que aparezcan duplicados aparentes que no son verdaderos.

Es por todo ello que las estimaciones que tradicionalmente se han venido haciendo a partir de los duplicados aparentes que hay en algunas grandes bases de datos estén siempre sesgadas y sistemáticamente agranden la cifra real de duplicados. Sólo puede evitar este sesgo una previa documentación rigurosa y exhaustiva de todos los duplicados observados, de modo que éstos sean contrastados con los documentos originales.

4. MÉTODO SENCILLO PARA CALCULAR APROXIMADAMENTE EL RUIDO EN ATRIBUTOS NO REPETIBLES DE UNA BASE DE DATOS

El ruido en los DNI de una base de datos aumenta de tal manera el número de duplicados internamente observados, que puede utilizarse este número (que es fácilmente medible) para obtener aquél.

Supongamos que una tabla de cierta base de datos contiene un único registro con DNI por cada persona de cierto colectivo. Llamemos D a la cantidad de DNI duplicados aparentes observados en dicha tabla, y llamemos r a la proporción buscada de DNI almacenados con errores. Sea N la cantidad de españoles vivos en posesión del DNI, $t = N/10^8$, y π la proporción de tarjetas del DNI en manos de españoles vivos cuyo número está duplicado. Sea n la cantidad de DNI que hay recogidos en la base. Supondremos que n es grande y que los errores en los DNI

producidos durante la recogida y almacenamiento de los mismos en la propia base de datos siguen una distribución uniforme.

Despejando r en la fórmula [1] del apartado 3.1:

$$r = ((N . D / n . (n-1)) - \pi) / (2 . (t - \pi))$$

$$\mathbf{r \approx 10^8 . D / (2 . n^2) - 0{,}0048}$$

Y, dentro de unos años, cuando ya no haya DNI duplicados:

$$\mathbf{r \approx 10^8 . D / (2 . n^2)}$$

Esta fórmula supone una vía sencilla, rápida y barata para obtener una primera aproximación del ruido en los DNI de una base de datos a partir de los duplicados observados internamente en la misma. Y es una contundente confirmación de que el ruido en las bases de datos va a ser permanentemente una fuente de generación de falsos duplicados.

Por otro lado, este procedimiento es fácilmente generalizable a cualquier base de datos que contenga algún atributo concebido para servir de identificador único de personas, objetos, situaciones, permisos y demás. Por ejemplo: matrículas de coche, números de afiliación a la Seguridad Social, números de identificación en el Censo Electoral, NIF, CIF, claves de licencias diversas, etc.

Para las Organizaciones responsables de estas bases de datos sería muy útil documentar sistemáticamente todos los errores puestos de manifiesto por éste o por cualquier otro procedimiento. Dicha documentación permitirá identificar las fuentes de incorrecciones y actuar sobre ellas, y también ayudará al experto a determinar el patrón de la distribución estadística de errores y a inferir con precisión el ruido existente en la base.

Obviamente, se podría refinar la fórmula anteriormente propuesta introduciendo otros factores, hipótesis, intervalos de confianza y demás, pero ello queda claramente fuera del ámbito de este estudio. Las enormes posibilidades de esta vía, así como su complejidad e interés práctico, bien merecen la atención futura de otros investigadores especializados.

5. AGRADECIMIENTOS

Este trabajo hubiera sido imposible de realizar sin la desinteresada colaboración de Blanca González, Mari Cruz Herranz, Inmaculada Muñoz, Ana Sánchez y Pilar Torrecilla.

Asimismo, fue del máximo interés el asesoramiento estadístico de Magdalena Cordero, Margarita González, Begoña Jáuregui y José Emilio Valdés.

Por último, se ha de citar el borrador de trabajo realizado por Ángel Gracia Guillén para el *Grupo Interministerial de Trabajo para la Unificación de Códigos de las Personas Físicas*, coordinado por el Consejo Superior de Informática. La inquietud manifestada en sus líneas provocó la motivación para hacer esta investigación, cuya realización fue posible gracias a la ayuda institucional facilitada por Frutos Abad.

6. BIBLIOGRAFÍA

Innumerables son los libros y revistas que tratan la teoría estadística usada en este estudio. Asimismo, son numerosos los artículos que mencionan la duplicación de los DNI, algunos de los cuales han sido incluso publicados por los *medios de masas*, dada la trascendencia social de este problema.

Sin embargo, todo parece indicar que no hay ninguna publicación estadística que trate el tema específico aquí desarrollado.

ESTIMATION OF THE DUPLICATES AMONG THE SPANISH NATIONAL IDENTITY DOCUMENTS (DNI)

SUMMARY

The code-number of the Spanish National Identity Document (DNI) is far from being a perfect database key, and many Spaniards have their code-numbers duplicated. Transcription mistakes are the most frequent cause of these duplicates, and they are very difficult to control since there are no evidence of them in the official registers. An estimation of the quantity of duplicated DNI code-numbers among Spaniards alive is presented, and a statistical model is offered to allow other estimates from different sources. The most important contribution is to have proved that the semiofficial estimates, inferred from data of huge databases not exempted of noise, were always skewed, magnifying the actual quantity of duplicates. Last, it is offered a procedure to estimate the database noise using the internal duplications of an unrepeatable identifier.

Key words: Spanish National Identity Document, DNI, duplication, transcription mistakes, database debugging.

AMS Classification: 62P99.