

Microdatos y k -anonimidad: un enfoque cuantitativo en el contexto español

Carlos J. Gil Bellosta

Datanalytics

cgb@datanalytics.com

16 de agosto de 2011

Resumen

Este estudio aporta una perspectiva cuantitativa al debate entre privacidad y transparencia en los datos públicos al trasladar al contexto español métodos y estimaciones realizadas en EE.UU. del número de personas que pueden ser identificadas en conjuntos de microdatos anonimizados pero en los que consta el sexo, fecha de nacimiento y lugar de residencia de los sujetos. Las estimaciones realizadas indican que el 42% de los ciudadanos españoles son reidentificables de conocerse tales características demográficas básicas.

1. Introducción

En aras de la transparencia y en reconocimiento del beneficio económico y social que aportan, muchos organismos públicos y privados divulgan microdatos de censos y encuestas. Por otro lado, la legislación de muchos países, entre ellos la española, recoge el derecho de los ciudadanos a la privacidad y el control sobre los datos personales que obran en poder de terceros. Estos derechos pueden verse comprometidos si la información contenida en algún fichero público de microdatos, aunque haya sido previamente anonimizado, permite identificar –ya sea inequívocamente o con un grado de certeza elevado– a un determinado ciudadano.

Esta tensión entre el derecho reconocido a la privacidad –que en este contexto equivale al anonimato– y el beneficio social de la difusión de datos –reconocido también en algunas jurisdicciones como obligación de las administraciones públicas– ha generado un debate al que este artículo quiere aportar algunas evidencias cuantitativas sobre hasta qué punto las salvaguardias mínimas que razonablemente y de oficio aplican las instituciones –como eliminar identificadores tales como nombres, apellidos, DNI, etc.– garantizan efectivamente la anonimidad de los ciudadanos.

El estudio está basado en dos trabajos, [9] y [2], que analizaron en el contexto estadounidense el siguiente problema: dadas determinadas características demográficas básicas de una persona, ¿en qué medida es posible

identificarla unívocamente? En el primero, usando datos del censo estadounidense de 1990, se llegó a la conclusión de que sería posible identificar al 87 % de la población sabidos su sexo, código postal y fecha de nacimiento. En el segundo, que utiliza como punto de partida el censo de 2000, se afirma que un 53 % de la población sería perfectamente identificable a partir de su sexo, municipio de residencia y fecha de nacimiento.

En este artículo se han trasladado estas cuestiones al ámbito español estimando el grado de anonimidad que cabe esperar en España en los microdatos de una encuesta o censo que recoja características demográficas básicas análogas. Para ello se han utilizado los datos del padrón municipal de 2010 y se han aplicado tanto los métodos de [2] como otros desarrollados por el autor.

La k -anonimidad o grado de anonimidad (véase [8]) en un conjunto de microdatos es un índice elemental que permite cuantificar hasta qué punto está garantizada la anonimidad de los individuos a los que se refiere: si en los microdatos consta un conjunto dado de características demográficas, se dice que un sujeto es k -anónimo si en el universo de la encuesta existen k individuos que comparten con él dichas características. Por ejemplo, si el sujeto es un varón de Ólvega nacido el 1 de agosto de 1960, entonces será k -anónimo si en dicho municipio viven k varones nacidos ese mismo día. Además, suele decirse que un conjunto de datos es k -anónimo cuando el grado de anonimidad de cada uno de los individuos que recoge es igual o mayor que k .

Obviamente, la 1-anonimidad equivale a una absoluta falta de anonimidad: un individuo 1-anónimo es perfectamente identificable. Aunque eso no signifique necesariamente que la reidentificación sea sencilla: aun conociendo –hipotéticamente– que solo hay un varón en Ólvega nacido el 1 de agosto de 1960, puede no resultar inmediato asociarle un nombre, un DNI, etc. No obstante, por ejemplo, una empresa que almacenase este tipo de datos demográficos de sus clientes y tuviese acceso a microdatos formalmente anónimos pero con un grado de anonimidad bajo podría desanonimizarlos y asociarles atributos de una manera, tal vez, ilegal. Esta posibilidad no es exclusivamente teórica: en [10] se describe el caso de un banquero estadounidense que cruzó datos de sus clientes con una lista de pacientes de cáncer y restringió el crédito a aquéllos que identificó como enfermos.

Hay que tener en cuenta también que aunque solo se consideran los datos sociodemográficos básicos a la hora de estudiar el grado de anonimidad, sería posible identificar unívocamente a un sujeto a partir de otros datos recogidos en una encuesta. Individuos con características particularmente infrecuentes –piénsese, por ejemplo, en niveles de renta muy elevados– pueden ser reidentificados incluso en ficheros de microdatos con un grado elevado de k -anonimidad. De ahí que se hayan elaborado medidas de anonimidad adicionales como la l -diversidad o la t -proximidad (véase [5]) que tienen en cuenta estas cuestiones. Cuestiones que, no obstante su importancia, no serán discutidas en este artículo.

También quedará fuera de su alcance la discusión de las técnicas y algoritmos que pueden utilizarse para enmascarar datos y aumentar el grado de anonimidad de los ficheros públicos. El lector interesado puede encontrar una discusión panorámica sobre dichas técnicas en [6].

2. Datos y métodos

Para el presente estudio se han utilizado los datos de distribución de la población española por sexo, municipio y grupos quinquenales de edad recopilados por el INE dentro de su explotación estadística del padrón del año 2010 [3]. El análisis se ha realizado usando el paquete de análisis estadístico R [7]. El código utilizado para realizar las estimaciones, gráficos y tablas que aparecen en el artículo está disponible en http://www.datanalytics.com/uploads/codigo_articulo_anonimidad.zip.

Para cada sexo y cada uno de los 8110 municipios españoles, las tablas del INE recogen el número de individuos con edades comprendidas entre 0 y 4 años, entre 5 y 9, etc. hasta el último tramo, que cuenta los mayores de 85 años de edad. Como no es posible conocer la fecha exacta de nacimiento de cada individuo, se ha supuesto que el día de nacimiento dentro de un intervalo quinquenal de edades sigue una distribución uniforme sobre los días del lustro que cubre. Se realizó la simplificación adicional de que la edad de los individuos del último grupo no excede los 90 años.

A pesar de que la distribución de los nacimientos a lo largo de año no es uniforme (véanse, por ejemplo, [4] y [1]) y de que existen personas de más de 90 años de edad, no se espera que estas simplificaciones desdibujen significativamente los resultados. Tampoco debieran alterarlos demasiado el considerar, como se ha hecho, todos los años y meses de igual duración al realizar estimaciones con fechas de nacimiento agrupadas en dichos niveles.

Supongamos que en cierto municipio residen n varones en un determinado grupo quinquenal de edad. Si N representa el número (entero) de periodos –días, meses o años– de un lustro, de acuerdo con [2], en promedio, existen

$$f_k(n) = \binom{n}{k} N^{1-n} (N-1)^{n-k} \quad (1)$$

periodos dentro de los N en los que han nacido k individuos. La demostración de este hecho es la siguiente: si X_i es el número de nacimientos habido en cada uno de los N periodos, el vector (X_1, \dots, X_N) sigue una distribución multinomial con vector de probabilidades asociado $(1/N, \dots, 1/N)$. La variable aleatoria que cuenta el número de periodos en el que ocurren k nacimientos es

$$\sum_{i=1}^N \mathbf{1}_{ik},$$

donde $\mathbf{1}_{i,k}$ es la variable aleatoria que toma los valores

$$\mathbf{1}_{ik} = \begin{cases} 1 & \text{si } X_i = k, \\ 0 & \text{si } X_i \neq k. \end{cases}$$

Su esperanza es

$$\mathbb{E} \left(\sum_{i=1}^N \mathbf{1}_{ik} \right) = \sum_{i=1}^N \mathbb{E} (\mathbf{1}_{ik}) = N \mathbb{E} (\mathbf{1}_{1k}) = N \mathbb{P}(X_1 = k),$$

y como cada X_i es una variable aleatoria binomial $B(n, 1/N)$, la expresión anterior es igual a

$$N \binom{n}{k} (1/N)^k (1 - 1/N)^{n-k}.$$

Reordenando los factores se obtiene, finalmente, (1).

Esta demostración, distinta de la que se ofrece en [2], permite generalizar el resultado al caso en el que las probabilidades de nacimiento en cada periodo son desiguales: si éstas fuesen p_1, \dots, p_N , la ecuación correspondiente a (1) sería

$$f_k(n) = \sum_i \binom{n}{k} p_i^k (1 - p_i)^{n-k}.$$

Retomando la discusión, de acuerdo con (1), habría un promedio de

$$k f_k(n) = k \binom{n}{k} N^{1-n} (N - 1)^{n-k}$$

varones k -anónimos en el intervalo de edad y municipio prefijados.

Por lo tanto, en el municipio completo, si n_{ij} representa el número de habitantes con sexo i e intervalo de edad j , en promedio se espera que haya

$$k \sum_{ij} \binom{n_{ij}}{k} N^{1-n_{ij}} (N - 1)^{n_{ij}-k}$$

individuos k -anónimos. Claramente, la suma puede extenderse a niveles de agregación distintos del municipal para obtener, por ejemplo, estimaciones provinciales o nacionales.

Una desventaja de este método es que, al calcular directamente el valor esperado del número de individuos k -anónimos, no permite medir la varianza de las estimaciones ni asociarles un intervalo de confianza. Por ese motivo, se ha acompañado de un método de estimación alternativo basado en simulaciones que permite aproximar la distribución del número de individuos k -anónimos en un ámbito determinado. Fijado un grado k de anonimidad, la simulación propuesta consiste en la iteración de los siguientes pasos:

1. Para iteración i y cada nivel de agregación (municipal, provincial o nacional) a , seleccionar los municipios m correspondientes y dentro de ellos, cada sexo s e intervalo de edad e .
2. Obtener el número de sujetos n_{mse} correspondiente a cada tupla (m, s, e) .
3. Obtener n_{mse} muestras aleatorias (con reemplazo) del conjunto $1, \dots, N$, que representan las fechas de nacimiento simuladas f de los habitantes de cada nivel de agregación.
4. Contar el número n_{ia} de tuplas (m, s, e, f) que se repiten exactamente k veces.

De esta manera, para cada nivel de agregación a , el conjunto de valores kn_{ia} , uno por cada iteración i , es una muestra de la distribución del estimador del número de individuos k anónimos dentro de a . Esto permite no sólo identificar el valor típico del estimador (habitualmente usando la mediana) sino también asignarle un intervalo de confianza. Los resultados de la simulación han puesto de manifiesto que dichos intervalos de confianza (véase el cuadro 1) son sumamente estrechos.

Es preciso que advertir que el algoritmo de simulación tal cual se ha descrito arriba es computacionalmente ineficiente. La implementación realizada en el código que acompaña al artículo, aunque fiel al esquema anterior, incorpora atajos para abreviar los cálculos.

k	día-mes-año			mes-año			año		
	%	n	±	%	n	±	%	n	±
1	42.38	1993	0.48	5.77	271	0.22	0.45	21	0.06
2	15.30	719	0.62	4.74	223	0.33	0.58	27	0.11
3	8.24	388	0.57	3.90	183	0.39	0.61	28	0.15
4	5.44	256	0.54	3.31	156	0.43	0.59	28	0.18
5	3.92	184	0.55	2.88	135	0.44	0.57	27	0.19
6-10	9.35	439	0.68	10.61	499	0.74	2.61	123	0.36
11-20	5.40	254	0.59	12.16	572	0.85	4.37	206	0.53
21-50	5.99	282	0.63	15.19	714	1.05	9.22	434	0.75
>50	3.98	187	0.52	41.44	1948	0.81	80.99	3808	0.63

Cuadro 1: Mediana del número de ciudadanos (en decenas de miles) empadronados en municipios españoles en función de su k -anonimidad junto con el intervalo de confianza al 95 % de la estimación en microdatos que contengan el municipio, el sexo y la fecha de nacimiento de los sujetos. Los tres bloques corresponden a los casos en que la fecha de nacimiento se especifica incluyendo el día exacto, el mes o únicamente el año.

3. Resultados

En esta sección se resumen los resultados obtenidos acerca del grado de anonimidad que cabe esperar en microdatos que contengan como información demográfica básica el municipio de residencia, el sexo y la fecha de nacimiento de los sujetos. El cuadro 1 muestra la distribución del número de ciudadanos en distintos niveles de k -anonimidad en función del grado de detalle con que se conozcan dichos atributos.

El 42,38 % (con un intervalo de confianza al 95 % de $\pm 0,01$ %) de los ciudadanos españoles, casi 20 millones de ellos, son perfectamente identificables de conocerse su municipio de residencia, sexo y la fecha de nacimiento. Eso convierte prácticamente a esta combinación de atributos en un *seudoidentificador*. Este porcentaje es equiparable en orden de magnitud a los obtenidos en [9] y [2] en el contexto estadounidense.

El porcentaje de ciudadanos reidentificables desciende al 5,77 % cuando de la fecha de nacimiento solo se conoce el mes y al 0,45 % cuando

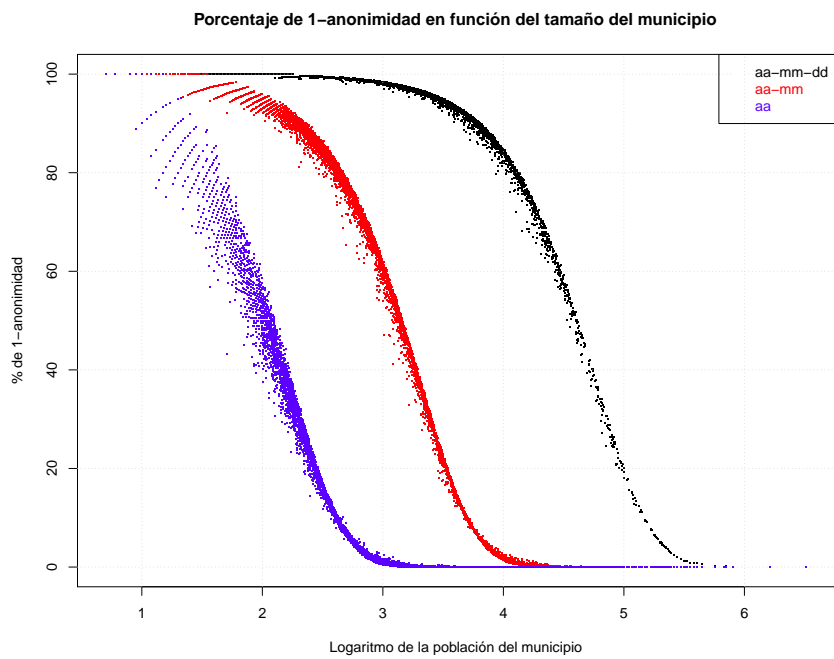


Figura 1: Porcentaje de población 1-anónima por municipio en función de su número de habitantes en conjuntos de microdatos en los que constan el sexo y la fecha de nacimiento de los sujetos. Los colores negro, rojo y azul corresponden a los casos en que de la fecha de nacimiento se conoce el día exacto, el mes o únicamente el año.

en élla solo consta el año. Este significativo aumento del grado de anonimidad subraya la efectividad de una técnica básica de anonimización, el de la confusión de atributos: dos fechas de nacimiento completas pueden *confundirse* cuando, por ejemplo, se elimina de ellas el día y quedan especificadas al nivel año-mes.

Merece la pena estudiar el grado de anonimidad a niveles distintos del nacional. Los cuadros 2, 3 y 4 muestran el número de habitantes esperado en varios grupos de k -anonimidad para una selección de municipios españoles. Por su parte, el gráfico 3 ilustra la relación entre el número de habitantes del municipio y el porcentaje de ellos que son 1-anónimos. Se observa cómo el principal determinante de la 1-anonimidad, habida cuenta de la distribución real –o habitual– de la población por sexos y grupos de edad, es esencialmente el número de habitantes de su municipio de residencia. Es notorio cómo es necesario que un municipio exceda los 100.000 habitantes para que le porcentaje de ciudadanos no anónimos baje del 20% y cómo en los municipios de menos de 10.000 habitantes más del 80% de los habitantes no pueden ser considerados anónimos de conocerse su fecha de nacimiento completa.

4. Conclusiones

Lograr un equilibrio entre transparencia y privacidad es un objetivo al que el presente estudio ha tratado de contribuir desde una perspectiva cuantitativa mostrando hasta qué punto las técnicas de anonimización más elementales pueden resultar ineficaces. Así, es costumbre de empresas y organismos públicos limitar el acceso a los datos de producción, que muchas veces contienen información sensible acerca de clientes y ciudadanos, pero a la vez poner en manos de conjunto mucho más amplio de empleados y usuarios datos –a la vista de los resultados de este estudio– solo superficialmente más anónimos borrando o enmascarando ciertos atributos.

El hecho de que casi la mitad de la ciudadanía española sea perfectamente reidentificable a partir de información demográfica básica subraya la relevancia la protección de datos y de la labor de las agencias oficiales encargadas de garantizarla. En efecto, velar por el derecho a la intimidad de los ciudadanos no puede limitarse a la aplicación rutinaria y pasiva de reglas básicas, tales como borrar campos identificativos; posee también una dimensión activa y con una importante componente analítica.

Referencias

- [1] Beresford, Geoffrey: *The Uniformity Assumption in the Birthday Problem*. *Mathematics Magazine*, 53(5):286–288, Noviembre 1980.
- [2] Golle, Philippe: *Revisiting the uniqueness of simple demographics in the US population*. En *Proceedings of the 5th ACM workshop on Privacy in electronic society*, WPES '06, páginas 77–80, New York, NY, USA, 2006. ACM, ISBN 1-59593-556-8. <http://dx.doi.org/10.1145/1179601.1179615>.

- [3] Instituto Nacional de Estadística: *Explotación estadística del padrón*. <http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft20%2Fe245&file=inebase>, 2010.
- [4] Lerchl, Alexander: *Where are the sunday babies? observations on a marked decline in weekend births in germany*. *Naturwissenschaften*, 92(12):592–4, 2005, ISSN 0028-1042. <http://www.biomedsearch.com/nih/Where-are-Sunday-babies-Observations/16205906.html>.
- [5] Li, Ninghui, Tiancheng Li y Suresh Venkatasubramanian: *t-closeness: Privacy beyond k-anonymity and l-diversity*. En *In ICDE*, 2007.
- [6] Matthews, Gregory y Ofer Harel: *Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy*. *Statistics Surveys*, 5:1–29, 2011. <http://www.i-journals.org/ss/viewarticle.php?id=74&layout=abstract>.
- [7] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. <http://www.R-project.org/>, ISBN 3-900051-07-0.
- [8] Samarati, Pierangela y Latanya Sweeney: *Generalizing data to provide anonymity when disclosing information (abstract)*. En *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, New York, NY, USA, 1998. ACM, ISBN 0-89791-996-3. <http://doi.acm.org/10.1145/275487.275508>.
- [9] Sweeney, Latanya: *Uniqueness of Simple Demographics in the U.S. Population*. Informe técnico, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.
- [10] Woodward, B.: *The computer-based patient record and confidentiality*. *N Engl J Med*, 333(21):1419–22, 1995.

municipio	grado de anonimidad										total
	1	2	3	4	5	10	20	50	100	>100	
Madrid	0	0	3	13	45	3371	39638	1381001	1832111	16867	3273049
Barcelona	6	48	193	529	1114	23279	357417	1214908	21840	3	1619337
Zaragoza	558	2274	5524	10706	18032	214367	389566	34095	0	0	675121
Bilbao	3261	11911	24554	37054	45513	189986	40888	21	0	0	353187
Alicante	4256	13370	25457	36870	44153	174250	36039	22	0	0	334418
Gijón	6354	18891	31760	40029	42260	125264	12640	1	0	0	277198
Oviedo	9332	24108	35421	39147	36208	77106	3833	0	0	0	225155
Tarrasa	10279	24772	35063	37466	33344	67713	4087	0	0	0	212724
Almería	11484	26026	35664	36784	30857	48163	1035	0	0	0	190013
Albacete	13368	27956	34900	32951	25548	35171	580	0	0	0	170475
Talavera de la Reina	21317	28374	21066	11212	4711	2303	1	0	0	0	88986
Benalmádena	21759	21187	11698	4696	1512	533	0	0	0	0	61383
Siero	22126	17977	8115	2647	685	179	0	0	0	1	51730
Molins de Rey	15839	6536	1547	271	39	5	0	0	0	0	24236
Unión (La)	13197	4284	774	100	10	1	0	0	0	0	18366
Pozo Alcón	4959	434	20	1	0	0	0	0	0	0	5413
Zalamea de la Serena	3678	231	8	0	0	0	0	0	0	0	3917
Selva	3406	207	7	0	0	0	0	0	0	0	3620
Guardiola de Berga	1007	17	0	0	0	0	0	0	0	0	1024
Torre la Ribera	117	0	0	0	0	0	0	0	0	0	117

Cuadro 2: Población esperada en una selección de municipios españoles en función de su grado de anonimidad de conocerse su sexo y fecha completa de nacimiento.

municipio	grado de anonimidad										total
	1	2	3	4	5	10	20	50	100	>100	
Madrid	0	0	0	0	0	0	0	0	0	3273049	3273049
Barcelona	0	0	0	0	0	0	0	0	0	1619337	1619337
Zaragoza	0	0	0	0	0	0	0	0	4876	670245	675121
Bilbao	0	0	0	0	0	0	0	1617	11303	340267	353187
Alicante	0	0	0	0	0	0	4	3195	22451	308768	334418
Gijón	0	0	0	0	0	0	0	2466	60515	214217	277198
Oviedo	0	0	0	0	0	0	28	4614	77547	142966	225155
Tarrasa	0	0	0	0	0	3	485	8689	76952	126595	212724
Almería	0	0	0	0	1	72	882	12161	74230	102667	190013
Albacete	0	0	0	0	0	31	988	16437	79970	73049	170475
Talavera de la Reina	0	1	3	10	23	411	2986	47069	38487	0	88986
Benalmádena	4	18	47	88	135	1249	5348	47709	6783	0	61383
Siero	1	5	15	33	56	855	12869	37634	264	0	51730
Molins de Rey	28	96	204	355	552	5990	12884	4125	0	0	24236
Unión (La)	86	221	377	542	715	6345	9583	496	0	0	18366
Pozo Alcón	438	1013	1283	1135	776	763	4	0	0	0	5413
Zalamea de la Serena	628	1091	1007	652	330	207	0	0	0	0	3917
Selva	651	1026	896	567	289	190	0	0	0	0	3620
Guardiola de Berga	624	304	79	15	2	0	0	0	0	0	1024
Torre la Ribera	111	6	0	0	0	0	0	0	0	0	117

Cuadro 3: Población esperada en una selección de municipios españoles en función de su grado de anonimidad de conocerse su sexo y mes de nacimiento.

municipio	grado de anonimidad										total
	1	2	3	4	5	10	20	50	100	>100	
Madrid	0	0	0	0	0	0	0	0	0	3273049	3273049
Barcelona	0	0	0	0	0	0	0	0	0	1619337	1619337
Zaragoza	0	0	0	0	0	0	0	0	0	675121	675121
Bilbao	0	0	0	0	0	0	0	0	0	353187	353187
Alicante	0	0	0	0	0	0	0	0	0	334418	334418
Gijón	0	0	0	0	0	0	0	0	0	277198	277198
Oviedo	0	0	0	0	0	0	0	0	0	225155	225155
Tarrasa	0	0	0	0	0	0	0	0	0	212724	212724
Almería	0	0	0	0	0	0	0	0	0	190013	190013
Albacete	0	0	0	0	0	0	0	0	0	170475	170475
Talavera de la Reina	0	0	0	0	0	0	0	0	182	88804	88986
Benalmádena	0	0	0	0	0	0	0	77	1020	60286	61383
Siero	0	0	0	0	0	0	0	0	370	51360	51730
Molins de Rey	0	0	0	0	0	0	7	332	3372	20525	24236
Unión (La)	0	0	0	0	0	17	73	995	3307	13974	18366
Pozo Alcón	0	1	2	4	5	35	303	5004	57	2	5413
Zalamea de la Serena	0	0	0	1	2	70	1175	2667	0	0	3917
Selva	0	0	1	4	7	134	1301	2177	0	0	3620
Guardiola de Berga	5	26	63	105	136	565	125	0	0	0	1024
Torre la Ribera	62	40	13	3	0	0	0	0	0	0	117

Cuadro 4: Población esperada en una selección de municipios españoles en función de su grado de anonimidad de conocerse su sexo y año de nacimiento.