ESTUDIO COMPARATIVO PARA DIFERENTES VALORES UMBRALES EN TABLAS DE FRECUENCIAS: UN EJEMPLO DE FUNCIONAMIENTO DE t-ARGUS.

Marta Mas

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

ESTUDIO COMPARATIVO PARA DIFERENTES VALORES UMBRALES EN TABLAS DE FRECUENCIAS: UN EJEMPLO DE FUNCIONAMIENTO DE t-ARGUS.

Marta Mas

RESUMEN

La elección de un valor umbral (n) apropiado para la protección de tablas de frecuencias es uno de los principales puntos a tener en cuenta por una agencia de estadística, a la hora de fijar una reglas genéricas para la preservación de la confidencialidad. Actualmente, diversos criterios son aplicados en los distintos institutos y oficinas de estadística.

Sin embargo, la decisión última sobre qué valor umbral es más adecuado aplicar, depende de los requerimientos de seguridad impuestos por la propia oficina de estadística. No obstante, se plantean algunos interrogantes a esta cuestión: ¿Cual es la mejor elección?, ¿Qué valor añadido se aporta a la calidad de la información si aumentamos este valor umbral?, ¿Vale la pena este incremento en términos de pérdida de información?.

En EUSTAT, se ha estado probando el software piloto τ-Argus, un programa especializado en la protección de macrotablas, y se han realizado diversas pruebas para distintos valores de los parámetros dentro de las reglas de sensibilidad que aplica el paquete. En el ejemplo de funcionamiento que se desarrolla a continuación, hemos utilizado dos macrotablas de frecuencias de la última encuesta del Censo en el País Vasco. Se han impuesto diferentes valores umbrales para cada tabla y se han observado las variaciones en el número de supresiones necesarias para proteger de forma adecuada cada una de las tablas propuestas. También se realizará una distinción entre supresiones primarias (celdas “inseguras”) y secundarias (protegen a las primarias), de forma que se puedan comparar sus incrementos.

Indice

Introducción

La elección de un valor umbral (n) apropiado para la protección de tablas de frecuencias es uno de los principales puntos a tener en cuenta por una agencia de estadística, a la hora de fijar una reglas genéricas para la preservación de la confidencialidad. Actualmente, diversos criterios son aplicados en los distintos institutos y oficinas de estadística. Los valores límite impuestos para las frecuencias dentro de cada celda varían desde 3 hasta 5 en algunos casos. A nivel nacional, la regla más extendida consiste en no publicar frecuencias inferiores a 3 cuando al menos una de las variables utilizadas en la definición de la tabla sea considerada como no-pública y los datos vayan referidos a un área geográfica de tamaño inferior a uno dado [1].

No obstante, la decisión última sobre qué valor umbral es más adecuado aplicar, depende de los requerimientos de seguridad impuestos por la propia oficina de estadística. Sin embargo se plantean algunos interrogantes a esta cuestión: ¿Cual es la mejor elección?, ¿Qué valor añadido se aporta a la calidad de la información si aumentamos este valor umbral?, ¿Vale la pena este incremento en términos de pérdida de información?.

En EUSTAT, se ha estado probando el software piloto τ-Argus [2], un programa especializado en la protección de macrotablas, y se han realizado diversas pruebas para distintos valores de los parámetros dentro de las reglas de sensibilidad que aplica el paquete. En el ejemplo de funcionamiento que se desarrolla a continuación, hemos utilizado dos macrotablas de frecuencias de la última encuesta del Censo en el País Vasco. Se han impuesto diferentes valores umbrales para cada tabla manteniendo constante, en cada caso, el valor de la *variable coste* (peso asignado a cada celda que determina su importancia en el procedimiento de supresión de celdas).

Se observará la variación del número total de supresiones que se necesitan para proteger de forma adecuada cada tabla. Es de esperar que dicho total crezca conforme aumenta el valor umbral para las celdas. También se realizará una distinción entre supresiones primarias (celdas “inseguras”) y secundarias (protegen a las primarias), de forma que se puedan comparar sus incrementos.

Descripción de variables y tablas

Los datos de entrada para el programa τ-Argus deben estar contenidos en un fichero ASCII de formato fijo. En este caso, partimos de un fichero donde cada registro o fila representa a un individuo y cada columna a una variable. Estos microdatos pertenecen a la encuesta del Censo realizada en el País Vasco en 1996. El número total de registros es de 2.098.055.

También es necesario definir un fichero que contenga información sobre las variables pero éste puede ser especificado de forma interactiva durante la sesión de trabajo en Argus. En un paso previo, se han seleccionado las variables que nos interesan y se han incluido en el fichero de entrada. Estas variables son las siguientes:

- *Lugar de Residencia*, 250 categorías (municipios).

- *Edad*, codificada originalmente por año pero agrupada posteriormente en tres categorías:

  - $\leq$ 19 años,

  - entre 20 y 64 años

  - $\geq$ 65 años

- *Situación de Residencia* dividida en dos categorías:

  - Residentes Presentes

  - Residentes Ausentes

- *Sexo*.

- *Tamaño del Municipio*, utilizada como variable coste.

La elección de la variable coste se basa en el modo en el que programa hace uso de la misma. El software permite a la persona que lleva a cabo la protección de los datos, guiar en cierto modo el procedimiento de supresión, asociando pesos a cada una de las celdas de la tabla. Estos pesos miden la importancia del valor de la celda por lo que a mayor valor de este peso menor probabilidad de que dicha celda sea suprimida. Si se considera la variable *Tamaño del Municipio* como variable coste, el procedimiento dará prioridad de supresión a aquellas celdas que representen valores para áreas geográficas pequeñas donde el riesgo de revelación de información confidencial es mayor. Esto supone una protección añadida a la tabla teniendo en cuenta el elevado nivel de desagregación geográfica que se considera en este caso (250 municipios).

Las tablas que se van a considerar para ser protegidas son las siguientes:

- *Lugar de Residencia* x *Edad* (tres grupos) x *Sexo*

- *Lugar de Residencia* x *Situación de Residencia* x *Sexo*

Ambas son tablas de frecuencias ya que las celdas representan recuentos de individuos y ambas generan un número relativamente pequeño de celdas inseguras para los valores umbrales aplicados comúnmente. Encontrar un patrón óptimo de supresiones para estas tablas en términos de pérdida de información, se convierte en un problema complejo de programación lineal. Argus puede resolver este problema en más o menos tiempo dependiendo fundamentalmente del número inicial de celdas sensibles a proteger. Es recomendable por lo tanto, lanzar el proceso de supresión de celdas partiendo del menor número de celdas inseguras posible. Este factor facilitará y agilizará el trabajo computaciones teniendo en cuenta que se van a chequear varios valores umbrales para cada una de las tablas.

Reglas de sensibilidad y valores de los parámetros

En el caso de tablas de magnitud Argus permite aplicar una regla de sensibilidad muy extendida basada en las contribuciones dominantes a la celda: la regla (N, p) [2]. Además, permite fijar un valor límite para la celda (n) que representa el mínimo numero de registros que deben contribuir a dicha celda, para que ésta no sea considerada insegura. Todas aquellas celdas que no cumplan alguno de estos dos criterios impuestos serán consideradas sensibles. El criterio aplicable para el caso de tablas de frecuencias que nos ocupa, será el del valor límite o umbral.

El objetivo consiste en aplicar diferentes valores límite a la misma tabla y comprobar el número de supresiones que generaría en cada caso. Es obvio que el rango de valores posibles para este valor umbral no es muy amplio. Si se considera $n \leq 1$, sólo las celdas unitarias serán marcadas como inseguras, lo cual no garantiza una protección suficiente en muchos casos. Por otro lado, si se impone un valor umbral de $n \leq 5$, por supuesto que no será fácil identificar a ningún individuo pero el coste que supondrá en términos de pérdida de información será mucho mayor. No obstante, cualquier valor entre 2 y 5 puede ser considerado aceptable pero, ¿qué se entiende por "aceptable"?. Será necesario encontrar un valor equilibrado de forma que aporte la protección requerida y preserve la mayor cantidad de información posible.

.

Ejemplo 1

Se ejecuta τ-Argus para la primera de las tablas propuestas: *Lugar de Residencia* x *Edad* x *Sexo* . La variable *Edad* que inicialmente es anual, ha sido recodificada en tres grupos antes del proceso de supresión. Para cada uno de los valores umbrales aplicados se atenderá al número de supresiones primarias (celdas sensibles o inseguras) y al de supresiones secundarias (aquellas que protegen a las primarias), una vez terminado el procedimiento de supresión de celdas llevado a cabo por el programa. La información generada por el programa sobre la ejecución de este proceso de supresión y los resultados obtenidos son almacenados en un fichero con los siguientes contenidos:

Produced 19:11:29 on 14/12/2000

/ Ficheros de entrada /

The input file with metadata is C:\Censo\Datos\Area1.rda (20:16:20 on 14/11/2000)

The input file with microdata is C:\Censo\Datos\Area1.asc (11:30:27 on 1/8/2000)

The table was saved in C:\Censo\Datos\Ejemplo1.ttb as follows:

Table:

MUNRv1 x EDAD3P x SEXOP : frequency ***/ Tabla de frecuencias /***

Cost variable for cells: TMUNR ***/ Variable Coste: Tamaño del municipio /***

The cell frequency limit 5 was applied; ***/ Valor Umbral n <= 5 /***

the safety range for each cell was [70%, 130%]. ***/ Rango de seguridad para cada celda /***

/ Supresiones /

There are **5** primary and **3** secondary suppressions in the elementary cells.

There are **0** primary and **6** secondary suppressions in the 2-dimensional marginals.

There are **0** primary and **0** secondary suppressions in the 1-dimensional marginals.

The general total was not suppressed.

"EDAD3P" has been recoded as follows: ***/ Recodificación de la variable Edad /***

1: 1- 19

2: 20- 64

3: 65-101

La información referida al número de supresiones generadas para cada uno de los valores umbrales aplicados en el *Ejemplo 1* ha sido resumida en la siguiente tabla:

Ejemplo 1.

Lugar de residencia x edad(3 grupos) x sexo		
Valor Umbral	**Tipo de supresión**	**Número de supresiones**
n ≤ 1	Primarias	0
	Secundarias	0
	Total	**0**
n ≤ 2	Primarias	0
	Secundarias	0
	Total	**0**
n ≤ 3	Primarias	1
	Secundarias	3
	Total	**4**
n ≤ 4	Primarias	3
	Secundarias	11
	Total	**14**
n ≤ 5	Primarias	5
	Secundarias	9
	Total	**14**
n ≤ 6	Primarias	8
	Secundarias	12
	Total	**20**

Como era de esperar, conforme crece el número de supresiones primarias más secundarias se necesitan para proteger a éstas. No obstante, se puede observar que las supresiones primarias en sí mismas suponen una protección adicional para la tabla. Es decir, para este ejemplo en el caso n ≤ 4, se necesitan 11 supresiones secundarias para proteger a 3 primarias, mientras que sólo son necesarias 9 secundarias para proteger a 5 primarias, en el caso n ≤ 5. Para este ejemplo, la elección del valor umbral bien sea 4 o 5 nos lleva al mismo resultado a efectos de número total de supresiones.

Ejemplo 2

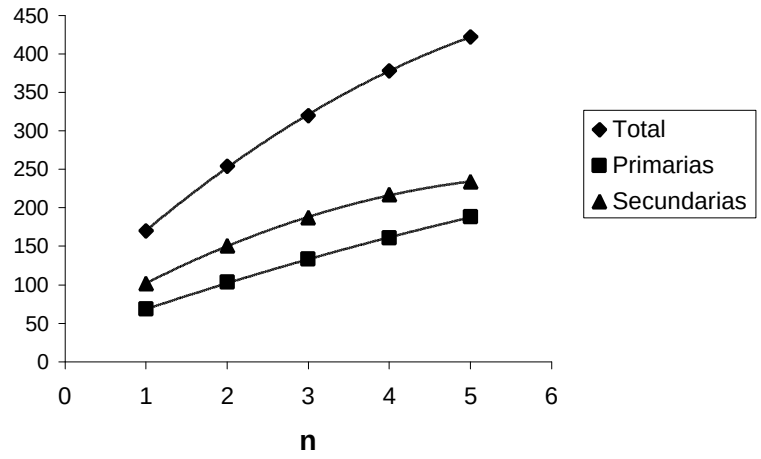
Se aplica el mismo procedimiento seguido para la primera tabla para este segundo ejemplo: *Lugar de Residencia* x *Situación de Residencia* x *Sexo*. Estos son los resultados obtenidos:

Ejemplo 2.

Lugar de residencia x *situación de residencia* x *sexo*		
Valor Umbral	**Tipo de supresión**	**Número de supresiones**
n ≤ 1	Primarias	68
	Secundarias	102
	Total	**170**
n ≤ 2	Primarias	103
	Secundarias	151
	Total	**254**
n ≤ 3	Primarias	133
	Secundarias	187
	Total	**320**
n ≤ 4	Primarias	161
	Secundarias	217
	Total	**378**
n ≤ 5	Primarias	188
	Secundarias	234
	Total	**422**

La tabla para este *Ejemplo 2* podría haber sido tratada mediante una recodificación previa de la variable *Situación de Residencia*. Para el caso n ≤ 1, esta característica genera todas las supresiones primarias en una única categoría especialmente sensitiva (*residentes ausentes*), por lo tanto se podría haber evitado la supresión de cualquier celda agregando previamente este grupo.

No obstante, el interés ahora radica en la observación de la variación del número de supresiones conforme varía el valor umbral aplicado. El siguiente gráfico muestra el crecimiento del número de supresiones para la tabla del *Ejemplo 2*:



El número de supresiones secundarias converge ligeramente hacia el número de supresiones primarias para valores "grandes" de n. Esta situación va a suavizar la línea de crecimiento que representa al total de supresiones necesarias para proteger la tabla.

Obviamente, el rango de posibles valores a tomar por el valor umbral no es muy amplio ya que no tiene mucho sentido, en la práctica, considerar valores para este límite mayores a 5. Sin embargo, puede ser interesante observar cómo se comporta el crecimiento del número total de supresiones conforme aumentamos el valor umbral y también el efecto que esto produce tanto en el número de supresiones primarias como secundarias.

Como hemos visto, en muchos casos se prefiere un mayor grado de protección asumiendo siempre un cierto coste añadido en términos de pérdida de información. No obstante, este coste no supone grandes incrementos si consideramos valores consecutivos del valor umbral impuesto. Por lo tanto, se puede aplicar un valor umbral lo suficientemente "sólido" como para aportar la protección requerida para la tabla y al mismo tiempo incentivar la confianza tanto de los encuestados como de los clientes de la estadística, con respecto a la preservación de la privacidad de la información individual.

Bibliografía

[1] GARÍN, A.. URRUTIA, J.

La preservación del secreto estadístico: elementos básicos de un sistema de protección de datos. Seminario OFISTAT. (Octubre 2000).

[2] C HUNDEPOOL, A.. WILLENBORG, L..

Tau-Argus. Versión 2.0 Manual del usuario. (Diciembre 1998).