Espacios Públicos
ISSN: 1665-8140
revista.espacios.publicos@gmail.com
Universidad Autónoma del Estado de México
México

Soberón Mora, José
La reconstrucción de bases de datos a partir de tablas de contingencias
Espacios Públicos, vol. 9, núm. 18, 2006, pp. 264-284
Universidad Autónoma del Estado de México
Toluca, México

Disponible en: http://www.redalyc.org/articulo.oa?id=67601819

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

La reconstrucción de bases de datos a partir de tablas de contingencias

José Soberón Mora[*]

Resumen

La técnica propuesta plantea la posibilidad de reconstruir la información contenida en las tablas de dos entradas utilizando la ponderación de datos. Consiste en usar los datos de cada una de las celdas, como ponderador en cada una de las intersecciones. Esto no genera la base de datos original, pero, como podrá observarse, es posible lograr procedimientos estadísticos más complejos que los obtenidos por el simple cuadro. Para lograrlo, se parte de ejemplos que tienen números absolutos o en su defecto aparecen las respectivas proporciones que pueden ser convertidas en absolutos con el fin de facilitar el cálculo de la nueva base. No obstante, el uso de este recurso tiene ya algunos años, su difusión ha tenido poco impacto entre los usuarios de máquinas computadoras y en el manejo de bases de datos. En este caso, se optó por la reconstrucción de las bases en SPSS por ser un paquete ampliamente conocido y comercializado.

Introducción

El procedimiento para construir una base de datos a partir de una tabla de análisis cruzado, también conocida como *tabla de contingencias*, *crosstab*, *tabla de doble entrada*, *tabla cruzada*, es poco novedoso. No obstante, todos los paquetes de análisis estadístico cuentan con la posibilidad de hacerlo, la recurrencia a este tipo de procedimientos es escasa debido en parte a que el usuario o el investigador, pocas veces se enfren-

[*] Maestro en Demografía por El Colegio de la Frontera Norte e investigador del Centro de Investigación y Estudios Avanzados de la Población de la UAEM.

ta a la necesidad de reconstruir una base de datos a partir únicamente de una tabla *crosstab*, o porque se desconoce la posibilidad de su construcción a partir de la herramienta propuesta en este documento. Esto es común cuando revisamos fuentes de datos necesarios para una investigación y se cuenta con información que resulta deficiente para nuestros fines o por haberse elaborado hace pocas décadas, donde el uso de los ordenadores era cosa de especialistas o sencillamente por la imposibilidad de contar con la base de datos original. Lo anterior también es común cuando nos topamos con información sobre una encuesta en particular y lo único que se ha localizado es un reporte de los resultados de la misma. Frecuentemente, el reporte localizado para el caso de las encuestas incluye tan sólo los cuadros de las variables de interés, si el objetivo de la encuesta era saber cuántos hijos había tenido por mujer, estas variables son reportadas en el único documento al que tenemos acceso.

La técnica de reconstrucción de datos a partir de tablas de cruce es de ayuda en la medida que no se desee reconstruir una base de datos a partir de ellas con la finalidad de tener la base original. Se debe tener presente, como se verá en el transcurso de la lectura, que esta herramienta produce lo mismo que se está observando en la tabla cruzada (y nada más). Entonces, en este momento, el lector se preguntará, ¿para qué molestarse en leer semejante documento como la herramienta que propone? La respuesta es clara. Una vez que se tiene la base reconstruida se pueden llevar a cabo procedimientos que de otra forma sería más que difícil porque necesitaríamos construir todos, y cada uno de los registros en forma de tabulado.

También se debe tener en cuenta que este procedimiento se propone para aquellos paquetes que manejen "archivo activo" por la verificación que se hace de la base de datos construida, no obstante, no existe limitación para llevarlo a cabo en cualquier programa. El SPSS, como muchos de los paquetes de estadística en el mercado, trabaja con una base de datos que se activó cuando se da la instrucción "abrir una base de datos". Los usuarios de otras versiones de este programa (versión 4 para PC o anteriores) como de otros programas, opcionalmente pueden escribir el procedimiento en un editor de textos para ejecutarlo como lote. Bajo esta modalidad, el usuario espera a que la computadora termine de realizar los cálculos respectivos y una vez finalizado procede a examinar el "archivo de resultados". La herramienta propuesta funciona en ambos casos.

PROCEDIMIENTO PARA LOGRAR UNA BASE *VIRTUAL*[1]

La reconstrucción de bases de datos requiere que se introduzcan los valores de una tabla (en lo sucesivo se llamará con este nombre a toda forma de presentación de datos que se citó en la introducción). Para lograr lo anterior se ha recurrido a dos ejemplos: Uno de ellos está presente en el manual de las primeras versiones de SPSS para PC relativo a "Correspondence Análisis" (ANACOR[2]) como material de referencia. El segundo, y más amplio de los ejem-

plos, es la base de datos del XXII Censo de Población y Vivienda 2000.

Ejemplo 1

En el manual respectivo de SPSS, encontramos una distribución hipotética con respecto a la orientación política de integrantes del parlamento europeo, de lo cual se reproduce lo que a continuación se presenta:

TABLA 1

ORIENTACIÓN DE CADA REPRESENTANTE

	Orientación			
	Demócrata Cristiano	Socialista	Otro	TOTAL
Bélgica	8	9	7	24
Alemania	39	30	6	75
Italia	25	11	39	75
Luxemburgo	3	2	1	6
Holanda	13	10	2	25
TOTAL	88	62	55	205

FUENTE: SPSS, 1992, p.33.

Esta tabla nos presenta dos situaciones:

1) El total de 205 casos nos indica que la base de datos original podía haber tenido 205 registros en total (cuando reconstruimos una base de datos bajo este procedimiento se debe tener presente que el total de registros pudo ser otro, normalmente mayor). Lo anterior obedece a la razón de que hemos eliminado sujetos que no respondieron a una de las dos variables analizadas en la tabla, este caso podría haber sucedido si un sujeto de Italia no hubiese respondido su tendencia política, tendríamos 206 registros en lugar de los 205, pero tendríamos un italiano con la opción "No respondió", la situación que acabo de describir es muy usual en encuestas y constantemente se debe tener presente la base de quienes respondieron a la pregunta. (Esto se verá con más detalle en el segundo ejemplo que utiliza el Censo mexicano de 2000).[3]

2) Nos brinda la posibilidad de conocer parte de lo que vamos a escribir en la sintaxis del SPSS.[4] El programa, ejecutable en la versión sintaxis hasta la versión 12, y que se encuentra disponible en el manual citado es:[5]

```
Ejemplo 1.1
data list free / nacion orientac factor.
begin data
        1   1    8   1   2    9   1   3    7
        2   1   39   2   2   30   2   3    6
        3   1   25   3   2   11   3   3   39
        4   1    3   4   2    2   4   3    1
        5   1   13   5   2   10   5   3    2
end data.
weight by factor.
```

```
Ejemplo 1.2

data list free / nacion orientac factor.
  variable labels
    nacion 'Pais representado'
    orientac 'Orientacion'
factor 'Factor de ponderacion' .
  value labels
     nacion 1 'Belgica' 2 'Alemania' 3 'Italia'
            4 'Luxemburgo' 5 'Holanda'/
     orientac 1 'Democrata cristiano'
              2 'Socialista' 3 'Otro'.
begin data
        1   1    8   1   2    9   1   3    7
        2   1   39   2   2   30   2   3    6
        3   1   25   3   2   11   3   3   39
        4   1    3   4   2    2   4   3    1
        5   1   13   5   2   10   5   3    2
end data.
weight by factor.
```

(Los ejemplos anteriores, el 1.1 y el 1.2, ejecutan lo mismo en SPSS, la única diferencia radica en que el primero es idéntico al que aparece en el manual respectivo y el segundo incluye darle la definición a las variables usadas. Para el caso de este ejemplo, se recomienda la aplicación de 1.2).

No obstante, se han definido los datos como una matriz de 9 x 5, el programa lo leerá como tres variables y lo ordenará como una base de datos con 15 casos y 3 variables, tal y como se definió en el "data list".

A esta nueva base de datos la llamaré base *virtual* porque únicamente contiene 15 registros, no obstante, no da la referencia de 205 casos. Podemos verificarlo si ejecutamos una frecuencia de cualquiera de las variables que definimos.

Veamos con más detalle lo que ha sucedido.

El ejemplo 1.1 indica que el caso 1 será leído como los sujetos bajo la condición 1 correspondiente a la nación que representa, Bélgica, también que pertenece a la agrupación de orientación demócrata cristiana y que en total son ocho sujetos, es por ello que la sección respectiva del programa dice:

begin data

1 1 8

El siguiente caso también pertenece a Bélgica, sólo que esta vez es socialista (un 2) y en la celda hay nueve sujetos, en la siguiente celda hay siete sujetos, luego 39 y así sucesivamente. La definición de cada una de las condiciones de las variables en este caso es 1 al 5 y 1 al 3 respectivamente, es indistinto el número que se utilice porque lo único que estamos definiendo son las características de una variable nominal. No sucede lo mismo con el número respectivo de cada celda porque éste nos indica el factor a ponderar o multiplicar. Lo que estamos haciendo es decirle al programa lo que vale cada una de las celdas hasta que en total tenemos 205 casos. En SPSS, la instrucción *weight by* efectúa la función de multiplicar el valor respectivo para cada uno de los casos. La misma instrucción se encuentra en todos los paquetes o programas que pueden manejar bases de datos de encuestas.

Cuando logramos introducir la base *virtual* a la computadora, existe la posibilidad de ejecutar cualquier operación estadística que nos permita el programa respectivo, por supuesto sin perder de vista el tipo de variables que estamos utilizando. En este momento es posible ejecutar desde una simple Chi cuadrada, hasta procedimientos más complicados como la regresión logística sin tener que utilizar la base de datos original. La limitante en este caso es la cantidad de variables que podemos utilizar, esto significa que este ejemplo dispone únicamente de dos variables.

Ejemplo 2

Lo que difiere de este ejemplo con respecto al uno es la presencia de la fuente de información. Primero se genera un cuadro a partir de la base de datos del Censo de 2000 para la entidad del Estado de México. De esta entidad se ha seleccionado el municipio de Acambay, que reporta 58 327 casos ponderados, de los cuales se utilizarán 51 144 porque son los casos válidos para ambas variables. El resultado queda de la siguiente forma (el lector puede ejecutar el crosstab respectivo y tendrá 51 144 casos válidos en este procedimiento. Si llegase a tener diferencia en el

número de casos puede atribuirse a la base de datos que proporciona la institución respectiva):

TABLA 2

RELACIÓN ENTRE NIVEL ACADÉMICO Y ASISTENCIA ACTUAL A LA ESCUELA
MUNICIPIO ACAMBAY

		Asistencia			
Código	Nivel académico	Sí va a la escuela	No va a la escuela	No especificado	Total
0	Ninguno	579	5248	76	5903
1	Preescolar o kinder	2095	132	35	2262
2	Primaria	10046	13681	88	23815
3	Secundaria	4220	7290	57	11567
4	Preparatoria o bachillerato	1600	1522	13	3135
5	Normal	19	153	0	172
6	Carrera técnica o comercial	117	783	0	900
7	Profesional	564	1120	1	1685
8	Maestría o doctorado	26	139	0	165
9	No especificado	91	1068	381	1540
Total sin ponderar		2059	3158	63	5280
Total ponderado		19357	31136	651	51144

Fuente: Elaboración propia con datos del XII Censo General de Población y Vivienda 2000.

Suponemos ahora que este cuadro es lo único de lo que se dispone y deseamos reconstruir la base de datos, el programa propuesto es:

```
data list free / nivel asiste factor.
      Value labels nivel   0 'Ninguno'
                           1 'Preescolar o kinder'
                           2 'Primaria'
                           3 'Secundaria'
                           4 'Preparatoria o bachillerato'
                           5 'Normal'
                           6 'Carrera técnica o comercial'
                           7 'Profesional'
                           8 'Maestria o doctorado'
                           9 'No especificado'/
                     asiste 1 'Sí va a la escuela'
                           2 'No va a la escuela'
                           3 'No especificado'.
```

```
begin data
 0  1  579   2  1  10046   4  1  1600   6  1  117   8  1  26
 0  2  5248  2  2  13681   4  2  1522   6  2  783   8  2  139
 0  3  76    2  3  88      4  3  13     6  3  0     8  3  0
 1  1  2095  3  1  4220    5  1  19     7  1  564   9  1  91
 1  2  132   3  2  7290    5  2  153    7  2  1120  9  2  1068
 1  3  35    3  3  57      5  3  0      7  3  1     9  3  381
end data.
weight by factor.
```

Al ejecutar el programa anterior, se forma una base de datos con 30 registros dentro del SPSS y, como en el ejemplo 1, la columna correspondiente a la variable factor indica el número de casos respectivos para cada una de las intersecciones de nuestra tabla.

¿QUÉ SUCEDIÓ?

En ambos ejemplos sencillamente se ha multiplicado cada una de las condiciones (celdas) por su respectiva frecuencia. Esto sucede con los 579 de la tabla 2 que en nivel académico mencionaron "ninguno" pero que sí van a la escuela.[6]

El usuario interesado puede reconstruir en una base de datos *virtual* el cuadro que desee, para ello necesita darle una definición de variable a cada una de las intersecciones (celdas) que intenta reconstruir.

TABLAS COMPLEMENTARIAS

La reconstrucción de datos de tablas complementarias es posible siempre que los totales sean complementarios, sin importar que los datos se encuentren en dos o más tablas separadas. Esto se puede observar en el siguiente ejemplo de las tablas 3.1 y 3.2, que contienen la información que ya se utilizó como ejemplo pero no estaba desagregada en hombres y mujeres. En este caso, la estructura de los datos es similar pero tenemos la información separada. La opción es captar cada una de las características como si fuera una sola y se propone que el programa quede de la siguiente forma:

```
data list free / nivel asiste sexo factor.
begin data
      0      1      1      335
      0      2      1      1369
      0      3      1      26
      1      1      1      1081
      .      .      .      .
      .      .      .      .
      8      1      2      0
      8      2      2      70
      8      3      2      0
      9      1      2      72
      9      2      2      748
      9      3      2      157
```

(Este ejemplo está incompleto y no puede ser utilizado como caso aplicable al SPSS)

La forma en la cual se captura la información de ambas tablas radica en que se debe añadir un 1 para el caso de los hombres y un 2 para las mujeres en el programa respectivo. El orden en el cual se ingresarán los datos de cada celda en el programa no afecta el resultado, no obstante, en la situación presentada, en las tablas 3.1 y 3.2 se opta por ingresar primero todas las celdas correspondientes a la tabla 3.1 para posteriormente hacerlo con la 3.2.

TABLA 3.1
RELACIÓN ENTRE NIVEL ACADÉMICO Y ASISTENCIA ACTUAL A LA ESCUELA (HOMBRES) MUNICIPIO ACAMBAY

		Asistencia			
		Sí va a la escuela	No va a la escuela	No especificado	Total
Código	Nivel académico				
0	Ninguno	335	1369	26	1730
1	Preescolar o kinder	1081	30	14	1125
2	Primaria	4792	6794	28	11614
3	Secundaria	2037	3788	26	5851
4	Preparatoria o bachillerato	844	783		1627
5	Normal	76	76		
6	Carrera técnica o comercial	44	157		201
7	Profesional	340	655	1	996
8	Maestría o doctorado	26	69		95
9	No especificado	19	320	224	563
Total sin ponderar		1008	1448	32	2488
Total ponderado		9518	14041	319	23878

FUENTE: Elaboración propia con datos del XII Censo General de Población y Vivienda 2000.

TABLA 3.2

RELACIÓN ENTRE NIVEL ACADÉMICO Y ASISTENCIA ACTUAL A LA ESCUELA (MUJERES) MUNICIPIO ACAMBAY

		Asistencia			
		Sí va a la escuela	No va a la escuela	No especificado	Total
Código	Nivel académico				
0	Ninguno	244	3879	50	4173
1	Preescolar o kinder	1014	102	21	1137
2	Primaria	5254	6887	60	12201
3	Secundaria	2183	3502	31	5716
4	Preparatoria o bachillerato	756	739	13	1508
5	Normal	19	77		96
6	Carrera técnica o comercial	73	626		699
7	Profesional	224	465		689
8	Maestría o doctorado		70		70
9	No especificado	72	748	157	977
Total sin ponderar		1051	1710	31	2712
Total ponderado		9839	17095	332	27266

FUENTE: Elaboración propia con datos del XII Censo General de Población y Vivienda 2000.

UTILIDAD

En una base de datos *virtual*, es posible la ejecución de todo tipo de operaciones tal y como lo hubiéramos efectuado con la base original, siempre y cuando se tengan presente las limitantes que se han mencionado, las cuales se agruparán en el apartado respectivo en el final de este documento.

Seguramente, el lector ya se dio cuenta que, con el ejemplo dos, se tiene la ventaja de reproducir directamente de una base de datos real, como es el censo del 2000. Los datos servirán para comprobar que la construcción del cuadro y la reconstrucción posterior de la base de datos son un recurso preciso.

Otra forma de mirar los resultados es por medio del análisis de correspondencias, que ya se citó anteriormente en este documento, en donde se examina la relación gráfica entre dos variables nominales.

Para ambos ejemplos lo que se generó fue una base de datos *virtual* a partir del contenido de una tabla. En el primero de los casos, el supuesto es que no contamos con la base de datos, en el segundo se recurrió a construir primero la tabla con una base disponible para posteriormente construir una *virtual* a partir del cuadro respectivo.

CONVERTIR PROPORCIONES

Seguramente, el interesado en reconstruir una base de datos, como la que se propone, puede verse en la necesidad de capturar proporciones como base de datos. La solución más sencilla es la conversión de proporciones en absolutos con la finalidad de llegar a tener absolutos dentro de cada celda. Uno de los objetivos radica en estimar un total de casos si no se cuenta con ello. Ésta es una circunstancia menos común pero ocurre cuando tenemos un cuadro con proporciones y sin total de sujetos.

El que no se nos informe sobre la cantidad de casos que se utilizaron para realizar el reporte, sucede con frecuencia en resultados de experimentos y las muestras no rebasan 30 casos. En estas circunstancias el cuadro respectivo ofrece poco.

Suponemos que la tabla 4 carece del total de casos con lo cual nos vemos impedidos a realizar la división respectiva. Lo primero que debemos realizar es conocer el número total de celdas que resulte de sumar 100%, esta opción, nos permitirá estar seguros que el total es por renglón, por columna o está calculado al total de la tabla. Continuemos con la intención de obtener un dato específico. El valor mínimo que podríamos tener de total de casos es 200 porque no es posible obtener una participación de 0.5% con un valor que no oscile en este valor o múltiplo del mismo. Posteriormente, si sabemos que el total de celdas es 15, multiplicamos 15 por .005 correspondiente a la celda con valor más bajo, obtendremos un valor de .075 casos, resultado poco convincente ya que esperaríamos un valor como mínimo de 1 por tratarse de un sujeto.

Como se puede observar, nos hemos referido a la intersección “Luxemburgo” con “Otra orientación”. El valor mínimo por el cual debemos multiplicar es 205 porque de otra forma, el decimal respectivo de cualquiera de los restantes 14 valores contendría un número que excedería el .1 al realizar la operación propuesta. Para lograrlo se ha recurrido a

reproducir la tabla en una hoja de cálculo y así conocer los valores que excederán el decimal citado.

En el primer caso, tenemos el mismo cuadro 1, pero en proporciones:

TABLA 4
ORIENTACIÓN DE CADA REPRESENTANTE

	Orientación			
	Demócrata Cristiano	Socialista	Otro	Total
Bélgica	3.9%	4.4%	3.4%	11.7%
Alemania	19.0%	14.6%	2.9%	36.6%
Italia	12.2%	5.4%	19.0%	36.6%
Luxemburgo	1.5%	1.0%	0.5%	2.9%
Holanda	6.3%	4.9%	1.0%	12.2%
Total	42.9%	30.2%	26.8%	100% (205 casos)

FUENTE: Tabla 1

¿Para qué deseamos conocer el total de casos?

Es evidente que para reconstruir la tabla, se requiere de un número de casos, el cual se aplicará como denominador en cada una de las celdas, trátese del total por columna, renglón o total del cuadro. En otro momento es posible reconstruir el total de sujetos que se emplearon en la elaboración de la tabla con la intención de asegurar que el procedimiento ha sido el adecuado principalmente cuando deseamos ejecutar el programa respectivo en un paquete estadístico.

Sugerencias

Como se observa en las tablas 3.1 y 3.2, se recomienda anotar ambos totales cuando se presentan los resultados de cualquier operación con una base de datos, esto con la finalidad de precisar al lector lo que lee, no obstante, algunos autores publican los datos ponderados, pero con los totales sin ponderar. Es cuestión de cada autor porque el poner cuadros más compactos reduce el espacio requerido en una página, además de asegurarse de que el posible lector no se pierda en explicaciones para saber si los resultados pertenecen al total ponderado o al total sin ponderar. En otra circunstancia se pudo eliminar la explicación de lo que hizo

sencillamente porque es obvio que los datos de una encuesta presentados en una tabla se deben ponderar.[7] Sobre esta situación existen ejemplos como la tabla presentada en la página 114 de Durand, *et al.* (2001) en donde se ha optado por la publicación de los porcentajes resultado de ponderar la base de datos pero con la muestra original "n", esto es 7 065 casos en total, resultado que nos daría 1 900 909 sujetos al ponderar.

Se optará por la opción que se desee para el manejo de los totales pero se sugiere tomar en consideración dos aspectos: 1) La presentación del total como de la tabla en general obedece objetivos específicos; 2) El presentar ambos totales permite al lector elaborar otros procedimientos como el factor de ponderación que se verá posteriormente.

TABLA 5

REGIÓN DE ORIGEN DE MEXICANOS QUE REPORTARON HABER TRABAJADO O BUSCADO TRABAJO EN LOS ESTADOS UNIDOS EN CUALQUIER MOMENTO DE SU VIDA. (LA FECHA CORRESPONDE A LA ÚLTIMA VEZ)

	Año de partida				
	1970-1974 (%)	1975-1979 (%)	1980-1984 (%)	1985-1989 (%)	1990-1992 (%)
Región de Origen					
Histórica	47.8	57.4	56.1	49.5	48.8
Fronteriza	33.7	26.3	27.4	24.8	29.7
Centro	17.0	13.8	15.2	24.6	19.8
Periferia	1.5	2.4	1.3	1.1	1.7
n =	489	813	1048	2276	2439

FUENTE: Durand, *et al*., 2001:114; con datos obtenidos de la Encuesta Nacional de la Dinámica Demográfica 1992.

El uso que se le dará a la base de datos *virtual* o reconstruida viene a responder parte de la pregunta formulada respecto a la utilidad del proceso. Un número de casos más elevado repercutirá en el valor de las significancias una vez que efectuemos cualquier prueba estadística.

RESPUESTA MÚLTIPLE

Cuando la suma de los respectivos porcentajes rebasa el 100%, es probable que estemos ante una tabla procedente de una pregunta con más de una posible respuesta. Si en una encuesta demográfica preguntamos la afiliación a los servicios de salud, la persona indicada nos puede responder que tiene el Instituto Mexicano del Seguro Social (IMSS) por

parte de su trabajo y el Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE) por parte de su pareja. La respuesta es correcta pero en la suma de las proporciones nos puede desconcertar mirar los resultados. El tratamiento para este tipo de preguntas es similar a lo que se ha sugerido en el caso de escalas nominales con la reserva de que la suma de las respectivas proporciones puede o no exceder 100%, dependiendo del número de casos, del número de decimales y de la cantidad de opciones en la respuesta múltiple.

En ocasiones únicamente quien construye la tabla, a partir de las variables utilizadas es la persona que sabe si la respuesta es múltiple o no. Lo anterior sucede con frecuencia porque la mayoría de los casos respondieron únicamente a la primera de las opciones en el cuestionario. Difícilmente, nos daríamos cuenta que una persona de entre mil tiene derecho a dos tipos de servicio médico. Por ello, se ha sugerido que las respectivas proporciones tengan cuando menos un decimal, con ello es más fácil enterarse si la respuesta fue o no múltiple.

De presentarse esta situación, es recomendable poner una nota aclaratoria, la cual indicará que la suma respectiva puede exceder el total esperado de 100%.

PROPORCIONES DE ESCALA

La herramienta propuesta en nada nos puede ayudar si lo único que tenemos a la mano es el promedio obtenido a partir de una variable numérica. En el caso de las encuestas, las escalas utilizadas permiten reconstruir las variables a partir de lo que respondieron los entrevistados, no obstante, que aparezca el promedio. En encuestas demográficas, el resultado podría presentarse como sigue:

TABLA 6
¿ALGUNA VEZ HAS ESTADO EMBARAZADA O HAS EMBARAZADO A ALGUIEN? (SÍ) ¿CUÁNTAS VECES?

	Sexo		Total
	Hombre	Mujer	
Una vez	46.6%	37.8%	41.0%
Dos	31.2%	33.7%	32.8%
Tres	11.6%	17.3%	15.2%
Cuatro	5.3%	6.1%	5.8%
Cinco	0.8%	2.0%	1.5%
Seis	0.2%	0.5%	0.4%

Siete	0.1%	0.2%	0.2%
Ocho	0.1%	0.1%	0.1%
Nueve	0.1%	0.0%	0.1%
Diez o más	0.1%	0.1%	0.1%
No especificado	4.1%	2.3%	2.9%
Promedio	1.8	2.2	1.9
Total sin ponderar	13186	17534	30720
Total ponderado	3472655	5945818	9418473

FUENTE: Cálculos propios con base de la Encuesta Nacional de Juventud 2000, realizada por el IMJ.

La tabla 6 muestra un valor de 0.0%, por el volumen de la base ponderada. Este dato únicamente significa que el valor es muy bajo y que aparecerá con 0. Ante esta situación es difícil conocer el valor ausente, no obstante, se darán algunas recomendaciones al final.

De manera similar, al tratamiento de las proporciones, en esta tabla se obtienen los valores de cada una de las celdas al multiplicar su porcentaje por el total ponderado. Una forma de verificar que la reconstrucción es correcta se logra al calcular el promedio en una hoja de cálculo, para ello se procede a multiplicar el correspondiente valor de la etiqueta (iniciaremos con el 1 porque indica un año) la cantidad que obtuvimos de multiplicar 3 472 655 x .466, la cantidad obtenida se suma con las demás y se divide entre el total ponderado.

LA PONDERACIÓN

Cuando se piense en ejecutar otros procedimientos estadísticos en una base *virtual* tales como la regresión logística, es posible reducir el número de casos en un cuadro ponderado, pero sin afectar las respectivas proporciones poblacionales.

La operación consiste en aplicar directamente la proporción respectiva a cada una de las celdas que integran nuestra base sea real o *virtual.*

Si la tabla cuenta con totales por variable, estos nos ayudarán a verificar que los casos asignados a cada celda se ajustan a lo esperado. Al retomar el ejemplo de la tabla 2, elaboramos la siguiente operación:

Dividimos la base de datos ponderada 51 144 entre la misma base pero sin ponderar 5 280 y el resultado lo aplicamos.

$$\frac{5\ 280}{51\ 144} = .103237916$$ (es importante incluir todos los decimales)

Este resultado lo multiplicamos por cada una de las celdas correspondiente, lo que nos da el resultado que deseamos porque la distribución proporcional se mantiene vigente si consideramos el ponderador original. Una manera de verificar que el procedimiento haya sido el adecuado consiste en obtener el total de sujetos no ponderados.[8]

A la operación antes descrita se le conoce como "coeficiente de expansión" o "factor de expansión, elevación o inflación". A la función inversa, de interés para nuestro trabajo, se le conoce como "*fracción de muestreo*" y normalmente se identifica con f (Cochran, 1971), pero en forma inversa, dado que deseamos regresar a los valores del total original. El coeficiente de expansión es:

$$\frac{\text{Total ponderado}}{\text{Total sin ponderar}} = \text{CE}$$

La fracción de muestreo es:

$$\frac{\text{Total sin ponderar}}{\text{Total ponderado}} = \text{f}$$

Lo que se ha sugerido sobre la ponderación tiene algunas implicaciones en el resultado. Específicamente, se recurrió a la ponderación relativa de la tabla 5 una vez que se dispuso de la base original de la encuesta que se utilizó para la construcción de la tabla, la Encuesta Nacional de la Dinámica Demográfica (ENADID) de 1992.

La fracción de muestreo utilizada fue 0.0037166429324076 porque es el resultado de dividir 1900 909 / 7065.

Al ejecutar el respectivo procedimiento para la obtención de la tabla (con el nuevo ponderador "relativo" en lugar del original),[9] se obtiene algo similar a lo presentado en este cuadro por las circunstancias siguientes. El total de 7 065 casos es el resultado original de la base de datos contenida en la respectiva sección de la ENADID,[10] pero al momento de ponderarla (con su ponderador original) el resultado es de un poco más de un millón novecientos mil casos. Al aplicar el ponderador

propuesto como fracción de muestreo es posible obtener el resultado en una sola ejecución del programa estadístico. La tabla 5 pudo haber sido construida originalmente con dos corridas de programa, una (sin ponderar) para conocer los totales de sujetos que daría cada columna 489, 813... y otra con el ponderador original de la base para conocer los respectivos porcentajes, "Región histórica 47.8%, 57.4%".

Dependiendo de las versiones de SPSS en la cual se obtenga el resultado, al aplicar el ponderador relativo a la base de datos para obtener la tabla 5, las versiones recientes convierten en "missing" tres de los sujetos originales. Esto sucede porque el programa está redondeando la cantidad de sujetos válidos para el análisis ejecutado. No obstante en la tabla de cruce respectivo únicamente aparecerá un total de 7 062 sujetos en lugar de 7 065. Únicamente en las versiones más recientes de SPSS,[11] aparece en el sumario de casos procesados la leyenda "Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded".

TENER DATOS DESAGREGADOS

El método tiene algunas particularidades si uno desea obtener los mayores beneficios al utilizarlo. Una de ellas tiene que ver con la forma en que se encuentran agregados los datos en las variables que deseamos reconstruir. En concreto, las variables deben mantener interacción, situación que se reflejará en la captación de datos como en las tablas 3.1 y 3.2, veamos otro ejemplo:

En la tabla 7, tenemos la presencia de tres variables que son orientación política, país de procedencia y sexo de la persona que ocupa la posición. Este tipo de relación nos presenta la interacción de dos variables, orientación política con país, y una de ellas no que en este caso se refiere al sexo del ocupante. Esto es claramente visible por el total que tenemos en la columna de la derecha que nos indica el mismo resultado en ambas secciones del cuadro.

Lo observable en este caso es que la variable sexo del ocupante no está relacionada con el cuadro como para formar una base *virtual* en donde intervengan las tres variables.

TABLA 7
Orientación

	Demócrata Cristiano	Socialista	Otro	Total
Bélgica	8	9	7	24
Alemania	39	30	6	75
Italia	25	11	39	75
Luxemburgo	3	2	1	6
Holanda	13	10	2	25

	Sexo		
	Hombres	Mujeres	Total
Bélgica	8	16	24
Alemania	30	45	75
Italia	22	53	75
Luxemburgo	4	2	6
Holanda	9	16	25
Total	73	130	205

FUENTE: Tabla 1 y datos hipotéticos.

No sucede lo mismo con la tabla 8, en donde cada una de las frecuencias estaría reflejando su interacción con las restantes. Si examinamos la primer casilla de esta tabla encontramos un 4 que significa que 4 sujetos del partido demócrata cristiano de Bélgica son hombres y 4 son mujeres que hacen un total de 8. En el caso de Alemania, el mismo partido tiene 20 y 19 respectivamente con un total de 39 sujetos.

Ambos ejemplos muestran la misma información, pero lo contenido en la tabla 8 permitirá captar toda la información de las tres variables a diferencia de la 7, en donde podríamos captar la relación para dos. Bajo este procedimiento, la variable sexo se tendría que captar en otra base de datos. Incluso, si se desea es posible incluir una columna del total de cada una de las intersecciones entre orientación política y país. De esta forma, se incluiría una columna de total que indicara los 25 sujetos italianos de orientación demócrata cristiana, tal y como lo indica la tabla 7.

TABLA 8
Orientación

	Demócrata Cristiano Sexo		Socialista Sexo		Otro Sexo		Total
	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	
Bélgica	4	4	5	4	4	3	24
Alemania	20	19	17	13	3	3	75
Italia	15	10	6	5	20	19	75
Luxemburgo	2	1	1	1	1		6
Holanda	8	5	6	4	1	1	25
Total	49	39	35	27	29	26	205

FUENTE: Tabla 1 y datos hipotéticos.

UTILIDAD DE LA PONDERACIÓN RELATIVA

He mencionado que usualmente los cuadros de cruzamiento que se reportan en los resultados de investigaciones o de encuestas reportan datos ponderados. De hecho, es lo que se debe reportar cuando existe un ponderador de por medio y se desea mostrar el dato encontrado, esto siempre sucederá, como ejemplo, con la base de datos del Censo de 2000. Surge entonces un pequeño problema al tener una cantidad elevada de registros y someterla a procedimientos como el análisis de regresión logística. El asunto es el siguiente, no se desea dejar sin ponderar la base de datos porque sabemos que los ponderadores pueden ser diferentes en cada unidad de muestreo, pero por otro lado tampoco se opta por ponderar la base porque prácticamente todo resultará significativo. Algunos usuarios lo que hacen es "relativizar" la base de datos. Esto se logra al dividir la base real sobre la base total (tal y como se ha realizado con anterioridad), el resultado será aplicado a cada registro de nuestra base *virtual* para obtener el número de casos original, pero sólo que en este caso la cantidad original, pero con su respectivo peso proporcional.

FIGURA 1
PASOS PARA LA CONSTRUCCIÓN DEL PONDERADOR RELATIVO

(1)	(2)	(3)	(4)	(5)
Contamos con la base original	se pondera para obtener el total de casos	dividimos el total de casos sin ponderar sobre el total ponderado	el resultado nos sirve como multiplicador para cada uno de los registros	Se obtiene la ponderación relativa a la proporción que representa
No contamos con la base original		dividimos el total de casos sin ponderar sobre el total ponderado	el resultado nos sirve como multiplicador para cada uno de los registros SIN DEJAR DE UTILIZAR EL PONDERADOR ORIGINAL	

Una solución práctica de lo anterior consiste en obtener un nuevo ponderador que sea el resultado de multiplicar el ponderador original por el obtenido en el paso 3 y 4 de la figura 1.

LO CONTRARIO (CON UNA BASE SIN PONDERAR, PONDERARLA)

Otra situación sucede cuando tenemos una base de datos sin ponderar y sabemos las proporciones reales de la población aludida. Resultaría raro que llegásemos a requerir auxiliarnos de un procedimiento similar, pero lo llevaremos a cabo. Imaginemos que tenemos la información del cuadro 1 en donde el total de los sujetos viene a ser 425, poco más del doble.

Procedimiento 1. Calcular las proporción respectiva de cada una de los cuadros para conocer el nuevo valor, en el cuadro 1 se observa que los 88 sujetos de orientación demócrata cristiana representan el 42.9% del total de sujetos, los 62 representan el 30.2% y los 55 el 26.8%. Como conocemos las proporciones reales de los sujetos tenemos que son 31%, 58% y 11% respectivamente. Lo único que debemos hacer es dividir la base de sujetos real sobre la ideal

RECOMENDACIONES

Al elaborar cuadros de doble entrada, se sugiere que se sigan estas recomendaciones independientemente de la información que se desee reportar en cada cuadro:

1. Colocar ambos totales al final del cuadro para conocer los dos valores y poder realizar la ponderación relativa en caso de requerirse. Los resultados al interior del cuadro estarán calculados de acuerdo con el total ponderado a reserva de indicar lo contrario.

2. Es conveniente que los valores reportados en proporción (%) contengan por lo menos un decimal. Algunos prefieren colocar en cada casilla ambos valores, el respectivo en absolutos y en proporción, pero en ocasiones se sacrifica estética y espacio del cuadro.

3. Todas las reconstrucciones de datos son susceptibles de ser verificadas al generar la misma tabla que capturamos. Los ejemplos de programa que se expusieron a lo largo del artículo deben verificarse para saber si cada una de las celdas contendrá el número adecuado de casos y para saber si la base total de casos es la correcta.

4. La tabla 6 muestra proporciones de 0.0%, por este motivo es recomendable ejecutar la misma tabla en absolutos para cerciorarse de que estas proporciones de valor 0 son valores que representan muy pocos casos, considérese lo que representarían en proporción uno o dos casos con un total de casi nueve millones y medio. Por otro lado, el agregar una nota aclaratoria a la respectiva tabla es de utilidad.

LIMITACIONES

Ya se mencionó que únicamente se podrán reconstruir las variables que estén en interacción ellas mismas. La mayoría de los reportes de investigación inclusive los censos de población presentan tablas de resultados o tabulados, de forma tal que son útiles para el usuario, pero hacen interactuar menos variables entre ellas.

NOTAS

[1] En lo sucesivo se llamará *virtual* a la base de datos lograda bajo este procedimiento.

[2] ANACOR = análisis de correspondencia.

[3] En la base de datos lee y procesa como *missing values* aquella ausencia de valor.

[4] Se ha elegido el uso de SPSS por las siguientes razones: es uno de los paquetes de estadística más comerciales que hay en el mercado y por su relativa facilidad de operación.

[5] El lector puede copiar directamente éste y los siguientes ejemplos (completos) de programa y ejecutarlo en cualquier versión, de la 4 a la 12, de SPSS.

Para este ejemplo, se ha utilizado uno aparecido en el manual respectivo de la versión 4.0.

[6] En el Censo de Población y Vivienda 2000, "Nivel Académico" se consideran los años concluidos de cada respondente.

[7] Esta afirmación es válida para el tipo de tablas como las que se señala. En otros procedimientos, la ponderación afecta a la elaboración de tablas como a la presentación de gráficas donde los datos reflejan años persona.

[8] Las proporciones originales no se pierden porque ya se aplicaron al momento de ejecutar la ponderación original. Esto sucede de la siguiente forma al construir un ejemplo ficticio: los migrantes internacionales con objeto de laborar viajan desde algunas regiones en una proporción de 80 hombres y 20 mujeres. Al levantar una encuesta, se podría realizar un muestreo de 50% y 50% respectivamente entre ambos sexos, para llegar a un total de 100 sujetos encuestados. Al momento de ponderar y proyectar la muestra se obtendría una base de 800 mujeres y 200 hombres porque se consideran las proporciones originales pero también a la población real. Al dividir el resultado ponderado por su factor de ponderación las proporciones respectivas se mantienen 80 y 20 porque lo único que se hace es disminuir el total de casos.

[9] Para obtener el nuevo ponderador se debe multiplicar el ponderador original en cada sujeto por la fracción de muestreo obtenida con la finalidad de conservar las proporciones originales.

[10] Se deben eliminar algunos sujetos como a los commuters.

[11] Se ha ejecutado en la versión 12.

BIBLIOGRAFÍA

Cochran, W. G. (1971), *Técnicas de muestreo*, México, CECSA.

Durand, Jorge, *et al*. (2001), "Mexican Immigration to the United States: Continuities and Changes" en *Latin American Research Review*, vol. 36, no.1, pp. 107-126.

Instituto Mexicano de la Juventud (IMJ) (2001), *Encuesta Nacional de Juventud 2000*, México, SEP-IMJ.

Instituto Nacional de Estadística, Geografía e Informática (INEGI), *Encuesta Nacional de la Dinámica Demográfica 1992*, *Base de datos*, México, INEGI.

Statistical Package for the Social Sciences (SPSS) (1992), *Manual Categories*, Chicago, SSPS.