

Universidad Católica del Uruguay
Facultad de Ciencias Humanas

Autotexto

Serie Estadística

Filtro y segmentación de archivos



**Universidad
Católica**

DAMASO A. LARRAÑAGA • URUGUAY

Laboratorio Metodológico

Versión original:

J. Bogliaccini – M. Cardoso – F. Rodríguez

Autores revisión 2008:

A. de León – M. Dodel – C. Rafaniello

Módulo de Práctica de Análisis

Serie Estadística

Tema: Selección de casos y segmentación de archivos

Descripción

En esta Guía veremos dos aplicaciones del SPSS que resultan muy útiles para el procesamiento de diversas bases de datos. La primera de ellas tiene que ver con la selección de determinados casos de la base que resultan de interés para el análisis que se quiere llevar a cabo. Por ejemplo, si quiero averiguar cuántos jóvenes de 12 a 17 años de edad asisten a los establecimientos de educación formal. Para ello, deberemos **seleccionar** a las personas que cumplan con el requisito de ser mayores de 11 años y menores de 18 y averiguar para ellas su asistencia a la educación. Ahora supongamos que nos interesa saber este mismo dato pero para Montevideo y para el Interior Urbano por separado. Hay varias alternativas para hacerlo, pero una de ellas es la segunda aplicación que veremos hoy: se puede **segmentar** el archivo por Montevideo e Interior Urbano y luego realizar el cálculo.

Usos de los comandos

Seleccionar casos

La selección de casos en una base de datos es una operación que se utiliza con mucha frecuencia para el análisis de la información. ¿Por qué? Básicamente porque permite aislar subpoblaciones de interés para el cálculo de determinados valores así como extraer una muestra de la totalidad de los casos. No olvidemos que normalmente las bases de datos contienen información de un conjunto muy amplio de observaciones, y muchas veces no nos interesa extraer información para todas ellas sino para un conjunto reducido de casos.

Segmentar el archivo

Esta opción es muy útil cuando se quiere hacer un análisis comparando dos o más conjuntos de poblaciones de nuestra base de datos. Por ejemplo, cuando se quieren obtener una serie larga de cuadros y frecuencias que comparen hombres y mujeres, basta con segmentar el archivo por dicha variable y realizar las distribuciones deseadas, sin necesidad de contemplar en el diseño del output la variable de segmentación.

Procedimiento

Las dos formas que veremos en esta Guía corresponden a las rutinas que se indican más adelante, las que permiten:

1. Seleccionar casos para realizar análisis de alguna población específica
2. Segmentar un archivo de datos para realizar análisis por grupos específicos de la población, comparando sus resultados.

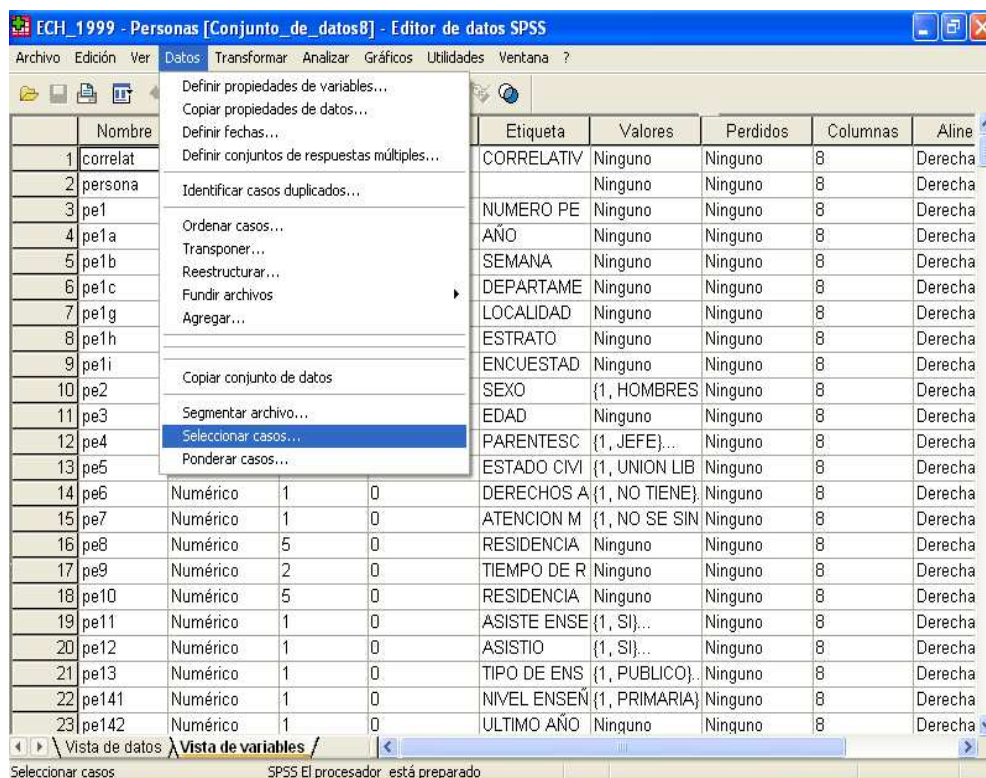
Ejemplo

Como ejemplo para esta Guía, utilizaremos la Encuesta Continua de Hogares (ECH) que realiza el Instituto Nacional de Estadística (INE). Dicha encuesta se practica en todo el país urbano de forma continua, es decir a lo largo de todo el año, desde hace más de 20 años consecutivos, proporcionando información para el cálculo de indicadores de la actividad laboral, de los ingresos de las personas y los hogares, sobre los años de educación alcanzados por las personas, etc. Para mayor información acerca de la metodología de dicha encuesta puede consultarse:

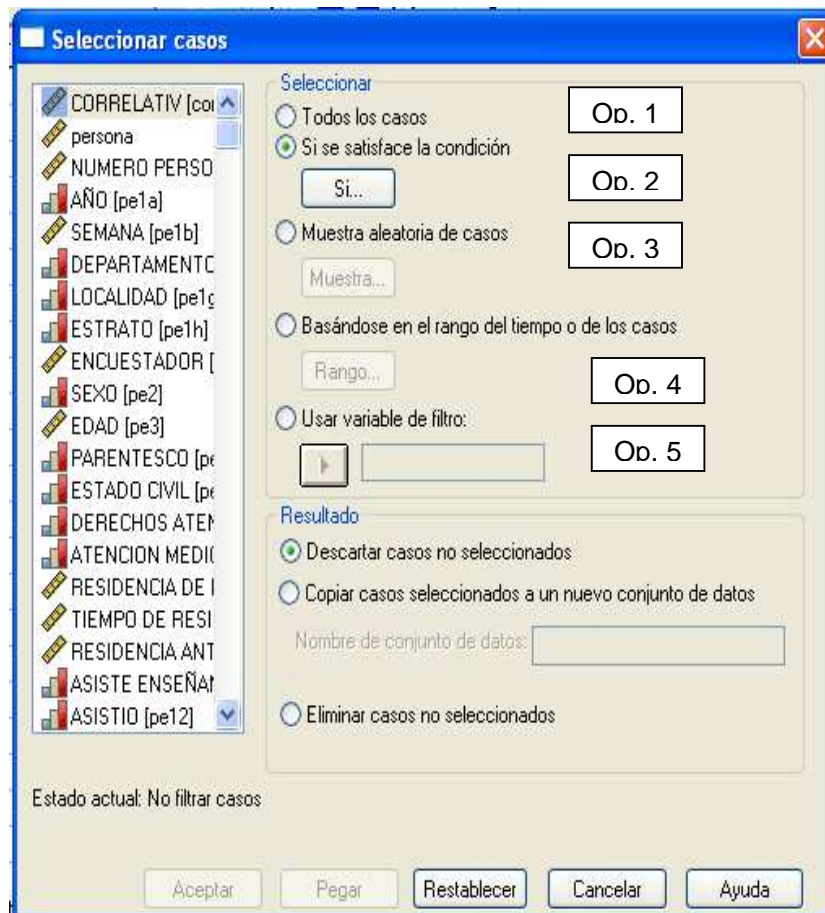
<http://www.ine.gub.uy/biblioteca/metodologias/ech/metodologiaech.htm>

Procedimiento para la Selección de Casos

El menú “Seleccionar casos” del SPSS ofrece varios métodos para seleccionar un subgrupo de ellos, basados en criterios que incluyen variables y expresiones complejas. La ruta de acceso es: *Datos / Seleccionar Casos*.

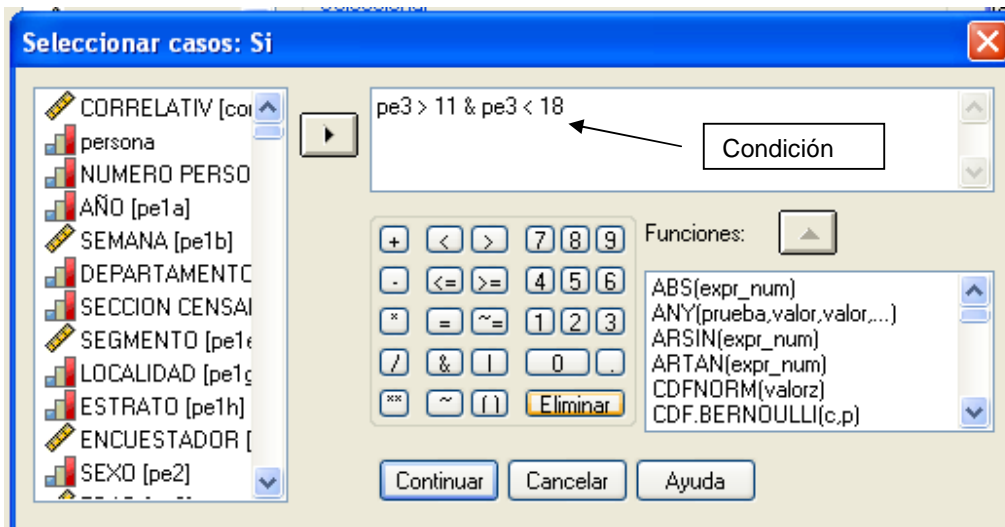


A partir de éste, se despliega un nuevo recuadro de diálogo con diversas opciones:

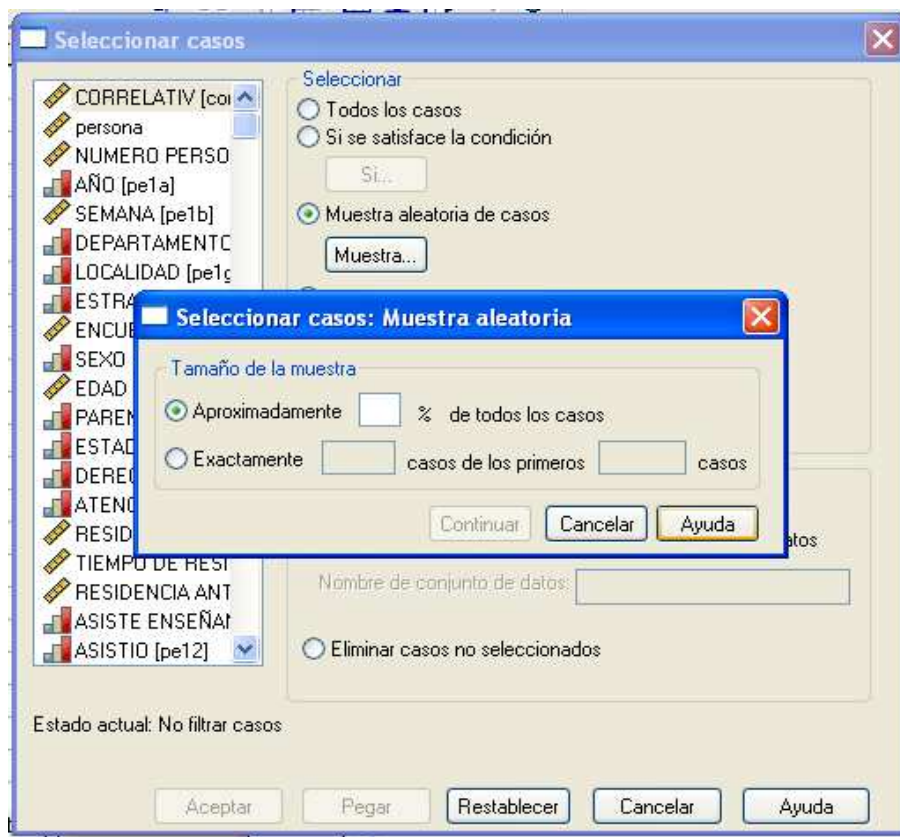


De no haber realizado un filtro de casos previo, este recuadro de diálogo aparecerá con la opción marcada en “*Todos los casos*”, es decir que tendremos seleccionados todos los casos de la base de datos (**Op. 1**).

La segunda opción (**Op. 2**) que ofrece este recuadro es la selección de casos que cumplan determinada condición. Esta opción es una de las más utilizadas ya que permite seleccionar los casos que verdaderamente nos interesa, por ejemplo la población de Montevideo. Para ello, debemos marcar el icono “*si satisface la condición*” y luego presionar el botón “Si...”, el cual nos habilitará un nuevo recuadro en el que deberemos poner el criterio de selección de los casos. Por ejemplo, si queremos realizar el análisis para la población de 12 a 17 años de edad, la condición deberá contemplar que la edad sea mayor que 11 y menor que 18.

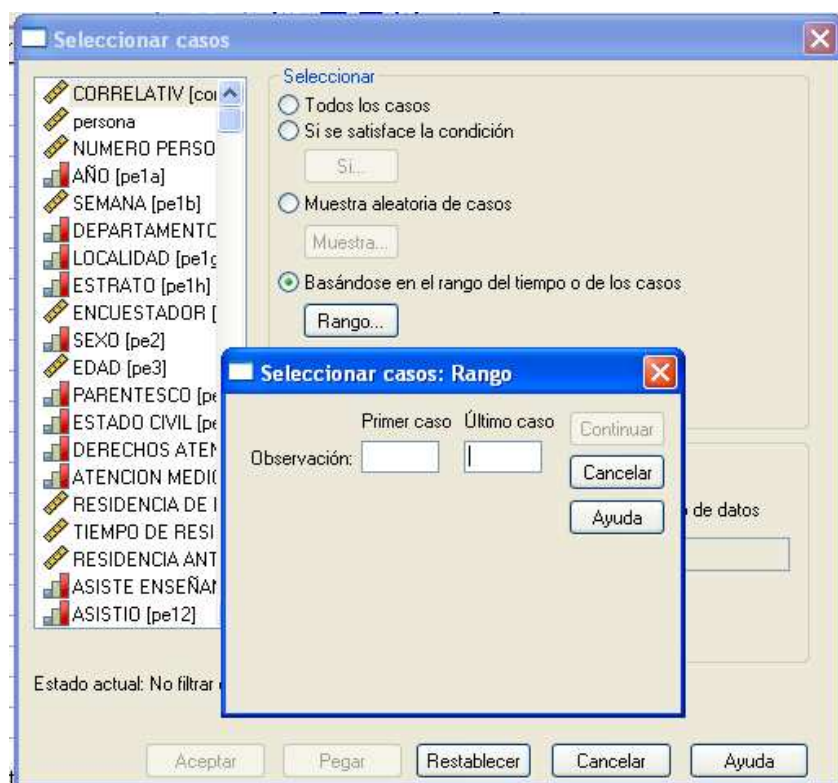


La siguiente opción (**Op. 3**) se denomina “*Muestra aleatoria de casos*”. Como lo dice su nombre permite seleccionar una muestra aleatoria y esto lo hace en base a un porcentaje aproximado o un número exacto de casos de los primeros X casos. La pantalla que se despliega es la siguiente:

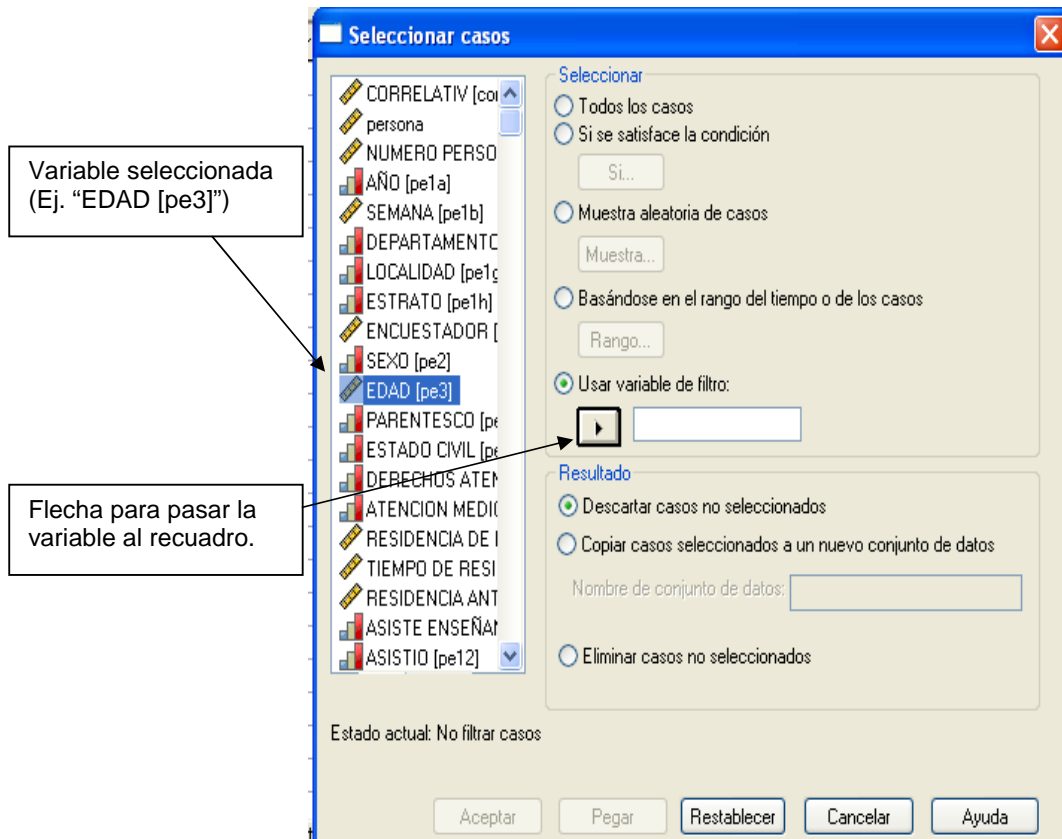


Luego de elegir entre una de esas dos formas de determinar la muestra deberá presionarse “Continuar” y luego “Aceptar”.

La cuarta opción (**Op. 4**) “Basándose en el rango de tiempo o de los casos” permitirá seleccionar determinada cantidad de casos correlativos. Por ejemplo, si se desea realizar un análisis de los primeros 100 casos se deberá utilizar esta opción para poder hacerlo. A partir de dicha selección, cualquier análisis que se lleve a cabo responderá únicamente a los 100 casos seleccionados. La pantalla que se despliega es la siguiente:



La quinta y última opción (**Op. 5**) “Usar variable de filtro”, es utilizada cuando se quiere practicar un análisis del archivo de datos sin considerar los casos que contienen “0” o no contienen dato (perdidos) en determinada variable. Cuando se requiera de hacer uso de este tipo de filtro únicamente se deberá seleccionar la variable de interés y colocarla en el recuadro desplegado. En este caso la variable elegida edad y para emplearla simplemente debemos seleccionarla y presionar la flecha que hay debajo de esa opción como se puede apreciar en la imagen.



Finalmente, se deberá marcar qué deseamos hacer con los **casos no seleccionados** en cualquiera de las 5 modalidades explicadas hasta el momento.

El programa nos plantea tres opciones: *“Descartar casos no seleccionados”* (lo que implica que los casos siguen estando en la base pero no son incluidos en los resultados de nuestros procesamientos); *“Copiar casos seleccionados a un nuevo conjunto de datos”* (se genera un nuevo archivo que únicamente contendrá los datos seleccionados y al cual podemos darle un nuevo nombre por medio de esta misma opción) y *“Eliminar casos no seleccionados”* (mediante esta alternativa borramos de la base en la que estamos trabajando los casos no seleccionados).

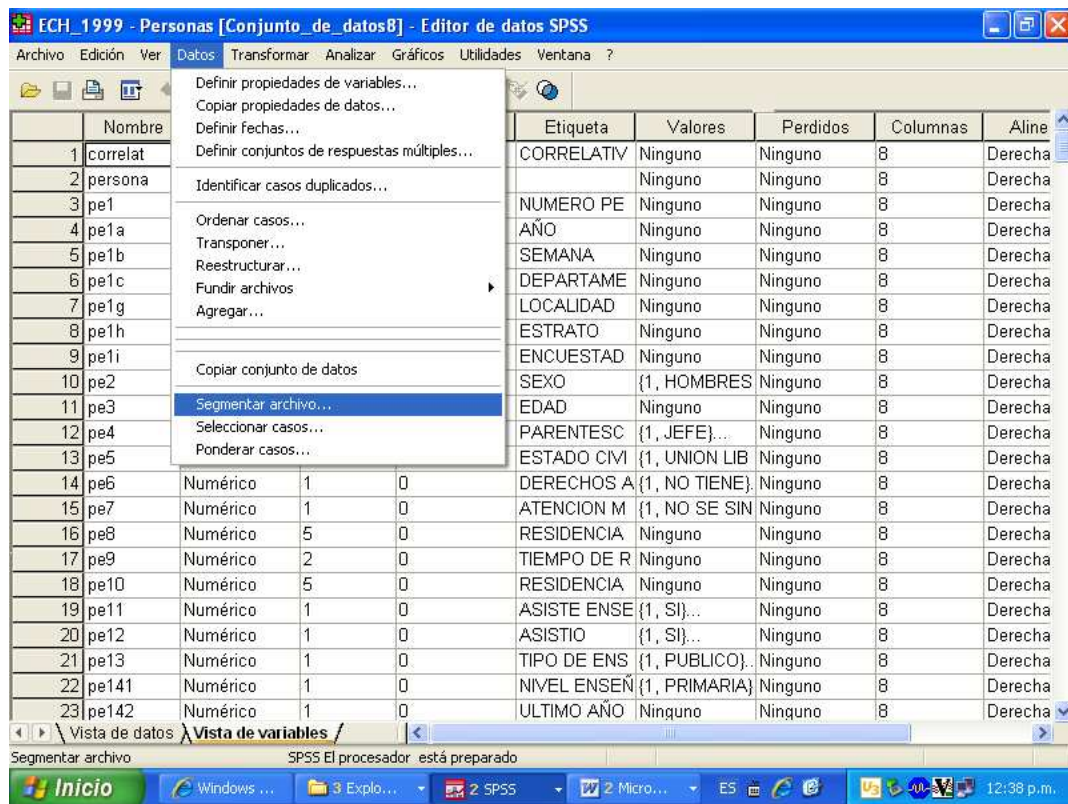
Recomendaciones

En lo que refiere a qué hacer con los casos no seleccionados debemos tener extremo cuidado. Por defecto aparece seleccionada la primera opción, *“Descartar casos no seleccionados”*, ya que es la más utilizada y, generalmente, es con ella que vamos a trabajar.

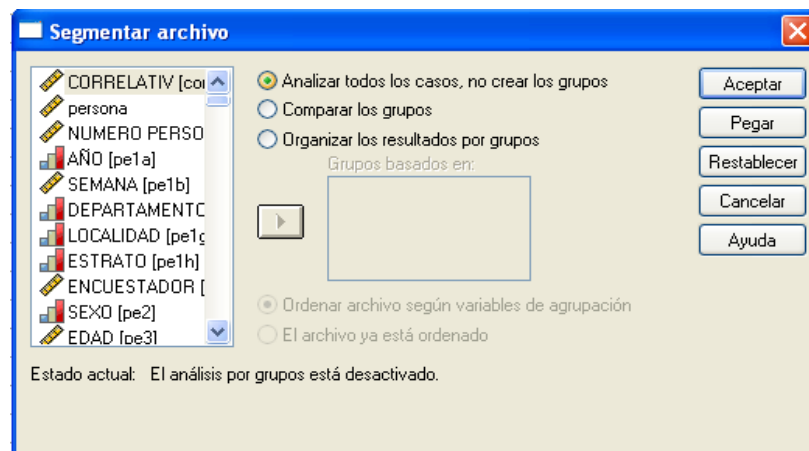
En caso de marcar la tercera opción, *“Eliminar casos no seleccionados”*, debemos tener la precaución de grabar el archivo con otro nombre para salvaguardar el original que contiene todos los casos. De no hacer esto perderíamos definitivamente los casos no seleccionados.

Procedimiento para la Segmentación de Archivos

Esta opción se encuentra en el menú *Datos / Segmentar Archivo*



Al hacerlo se nos despliega la siguiente ventana:



La primera opción, por defecto, no realiza comparación alguna. Si queremos hacerla se nos brinda la posibilidad a través de dos formas: "Comparar los grupos", y "Organizar los resultados por grupos". La diferencia entre ambas es únicamente visual, en la forma de presentar las salidas. Mientras que la primera lo hace en la misma tabla la segunda genera una salida para cada categoría de la variable de segmentación.

Por ejemplo, si segmentamos por género y sacamos una frecuencia del estado civil, en la primera opción se obtiene lo siguiente:

PARENTESCO

SEXO	Frequency	Percent	Valid Percent	Cumulative Percent
HOMBRES Valid	JEFE	20	46.5	46.5
	CONYUGE	1	2.3	48.8
	HIJO DE AMBOS	9	20.9	69.8
	HIJO SOLO DEL JEFE	9	20.9	90.7
	HIJO SOLO DEL CONYUGE	1	2.3	93.0
	OTRO FAMILIAR	1	2.3	95.3
	NO PARIENTE	2	4.7	100.0
	Total	43	100.0	100.0
MUJERES Valid	JEFE	17	29.8	29.8
	CONYUGE	16	28.1	57.9
	HIJO DE AMBOS	11	19.3	77.2
	HIJO SOLO DEL JEFE	5	8.8	86.0
	YERNO / NUERA	1	1.8	87.7
	NIETO/A	1	1.8	89.5
	PADRES-SUEGROS	1	1.8	91.2
	OTRO FAMILIAR	1	1.8	93.0
	NO PARIENTE	4	7.0	100.0
	Total	57	100.0	100.0

mientras que en la segunda es esta otra:

PARENTESCO

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid JEFE	17	29.8	29.8	29.8
CONYUGE	16	28.1	28.1	57.9
HIJO DE AMBOS	11	19.3	19.3	77.2
HIJO SOLO DEL JEFE	5	8.8	8.8	86.0
YERNO / NUERA	1	1.8	1.8	87.7
NIETO/A	1	1.8	1.8	89.5
PADRES-SUEGROS	1	1.8	1.8	91.2
OTRO FAMILIAR	1	1.8	1.8	93.0
NO PARIENTE	4	7.0	7.0	100.0
Total	57	100.0	100.0	

a. SEXO = MUJERES

PARENTESCO

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid JEFE	20	46.5	46.5	46.5
CONYUGE	1	2.3	2.3	48.8
HIJO DE AMBOS	9	20.9	20.9	69.8
HIJO SOLO DEL JEFE	9	20.9	20.9	90.7
HIJO SOLO DEL CONYUGE	1	2.3	2.3	93.0
OTRO FAMILIAR	1	2.3	2.3	95.3
NO PARIENTE	2	4.7	4.7	100.0
Total	43	100.0	100.0	

a. SEXO = HOMBRES

Ejercicios

Ejercicio A

- a) Utilizando la base de datos de Encuesta Continua de Hogares seleccione la población menor de 40 años de edad y realice una frecuencia del parentesco de las personas con el jefe del hogar, comparando los hombres con las mujeres.
- b) La categoría “Jefe del Hogar”, se distribuye igual según el sexo? ¿Y la del cónyuge?
- c) ¿Qué sucede con los hijos en relación al sexo?

Ejercicio B

- a) Utilizando la base de datos de Encuesta Continua de Hogares seleccione la población de Montevideo, menor de 40 años de edad y realice una frecuencia del parentesco de las personas con el jefe del hogar, comparando los hombres con las mujeres.
- b) ¿Qué diferencias encuentra entre esta distribución y la del conjunto del país realizada en el Ejercicio A?
- c) Ahora pruebe hacer el mismo análisis, pero utilizando la opción *Segmentar archivo* en lugar de *Seleccionar casos*.

Bibliografía de referencia

- Blalock, H. (1966) Estadística Social. FCE. México.
 - Peña, D. Romo, J - Introducción a la Estadística para las Ciencias Sociales. Mc Graw Hill 1997.
 - Mason y Lind - Estadística para administración y economía. Alfaomega 1998. México, D.F. 8ª edición.
-