MANUAL BÁSICO DE ANÁLISIS DE DATOS CON PSPP

Juan Moncada Herrera
Departamento de Ciencias Matemáticas y Física
Universidad Católica de Temuco
jmoncada@uctemuco.cl

CONCEPTOS PRELIMINARES

El análisis de datos depende directamente de la naturaleza de los mismos. Por tal razón, una de las primeras cuestiones a resolver a la hora de intentar un análisis de datos tiene que ver precisamente con el tipo de datos disponible. En este sentido, hay sólo dos alternativas para la naturaleza de la información (o de los datos): Información o datos *cualitativos*, o bien información o datos *cuantitativos*.

Si la información disponible es de naturaleza cualitativa, y dependiendo del tipo de investigación, los análisis pueden consistir de: Análisis narrativo, Semiótica, Análisis de contenido, Análisis del discurso, Teoría fundamentada, Hermenéutica, Fenomenología, entre otras formas de análisis. A su vez, los procedimientos más frecuentemente utilizados son codificación, inducción analítica, mapas cognitivos, incidentes críticos, entre otros. En cambio, si la información que se desea analizar es de tipo cuantitativo, entonces su análisis suele ser de tipo estadístico, entendiendo por tal la movilización de una serie de recursos y procedimientos tendientes a la elaboración de síntesis de la información con el fin de proveer de mensajes lo más informativos posible. Al igual que en el caso cualitativo, estos análisis naturalmente dependen de los objetivos que se persiguen en la investigación. De modo muy general puede tratarse de análisis descriptivos, análisis comparativos, análisis relacionales, análisis para la reducción del número de variables, análisis para la agrupación y la clasificación de elementos, entre otros.

La Tabla de Datos

Un aspecto fundamental en el análisis estadístico de los datos es su disposición en una estructura determinada. Esta estructura es generalmente recogida en lo que se denomina *Tabla de Datos*. Así como los objetivos son uno de los aspectos más relevantes en una investigación, la Tabla de Datos es también la componente más importante en el proceso de análisis. Tan importante es, que si no se dispone de ella no es posible el análisis (estadístico) de los datos.

La construcción de la tabla de datos es un proceso que se inicia en la formulación misma de la investigación, y debe orientar, de manera especial, la elaboración de los instrumentos de recogida de información. Construir una tabla de datos significa adelantarse a las categorías que la investigación revelará, codificar y recodificar información a priori o a posteriori, precisar la naturaleza de la información que se obtendrá así como las unidades sobre las que se llevarán a cabo las distintas mediciones o sobre las que se observarán características

previamente definidas. Como puede apreciarse, la elaboración de la tabla de datos puede ser un proceso complejo, que requiere generalmente de la participación de equipos o de expertos en la temática en estudio.

Una Tabla de Datos no es más ni menos que una disposición de la forma siguiente:

Individuo	**Variable 1**	**Variable 2**	…	**Variable *p***
1	Valor 1_1	Valor 1_2	…	Valor 1_*p*
2	Valor 2_1	Valor 2_2	…	Valor 2_*p*
…	…	…	…	…
N	Valor *n*_1	Valor *n*_2	…	Valor *n*_*p*

Estructura de una Tabla de Datos

En una investigación cada uno de los elementos que constituyen la tabla de datos debe estar claramente definido. Esto que, a primera vista parece bastante obvio, a veces resulta no serlo tanto. Por ejemplo, la Encuesta CASEN recoge diversa y variada información socio-demográfica por medio de un instrumento que contiene una serie de preguntas que el entrevistado debe responder, ya sea a título personal, o en representación de los integrantes de su grupo familiar. En este contexto surgen tres unidades de observación (o individuos estadísticos): el o la Jefe de Hogar, cada integrante del grupo familiar, y la propia unidad familiar (la vivienda). Así, la definición, reconocimiento y precisión de los individuos estadísticos y de las variables correspondientes pasan a formar parte de uno de los procesos más relevantes en la estructura de cualquier encuesta.

Por todo lo anterior, la tabla de datos es un elemento al que se debe prestar la mayor atención posible, después del proceso de recogida de información.

Recursos informáticos para el análisis de datos

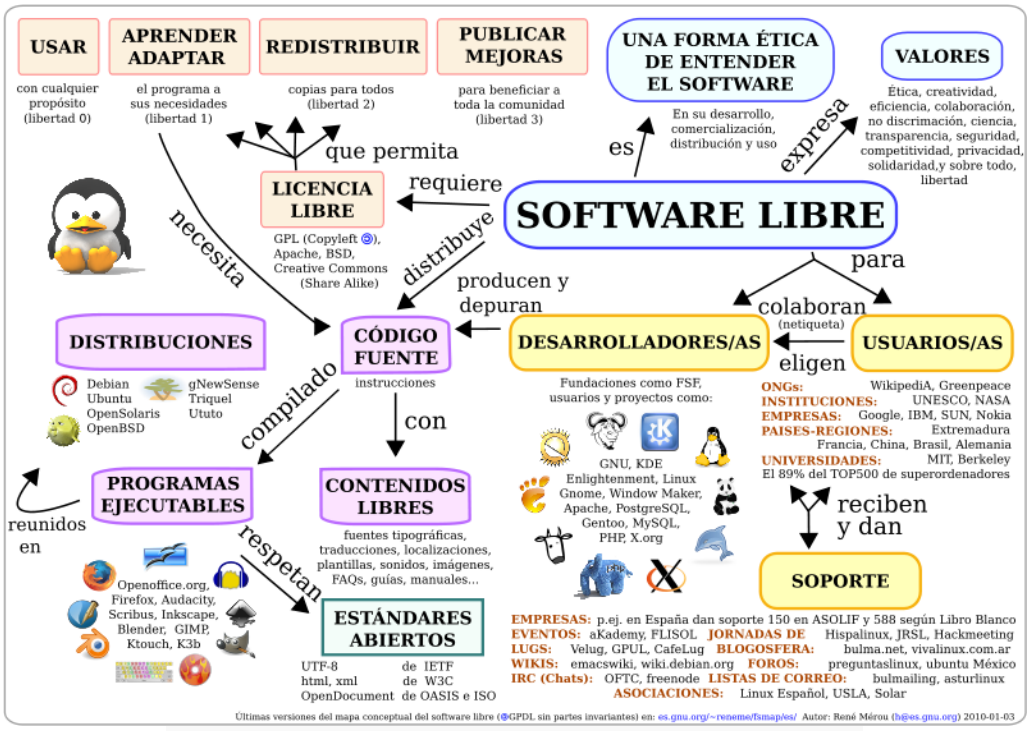
En muchas situaciones, y particularmente en la actualidad, el análisis de la información o el procesamiento de los datos es casi imposible realizarlos con herramientas elementales y básicas, como una calculadora. Esto debido a los grandes volúmenes de información disponibles, o simplemente porque estándares internacionales y la propia práctica científica requieren que los análisis estadísticos sean llevados a cabo con ayuda de herramientas y recursos que den garantía de precisión y validez. En estos casos se hace necesario el uso de sistemas especializados en el tratamiento de la información, como los provenientes de la Informática y la Computación. La primera disponiendo de programas que facilitan los cálculos y la organización de datos, y la segunda proporcionando instrumentos técnicos que facilitan el trabajo de los primeros.

En la actualidad es casi inimaginable el tratamiento estadístico sin el apoyo de software estadístico. Entonces surge un problema adicional que resolver a la hora de decidir tratar estadísticamente un conjunto de datos: ¿qué software utilizar? La respuesta a veces no es tan simple. Requiere de conocer el tipo o naturaleza de los datos, las preguntas que se desee responder a partir de los datos, la estructura en la que se nos presentan los datos, etc. Hoy en día existe una infinidad de productos informáticos orientados al tratamiento estadístico de datos. Los hay libres y gratuitos, cuya filosofía es el crecimiento y desarrollo colectivo de la sociedad, y cuyo desarrollo es producto del esfuerzo de grupos y equipos humanos que trabajan, muchas veces de forma desinteresada, porque otros progresen y hagan progresar a la sociedad a la cual pertenecen. En esta línea de trabajo se pueden encontrar desde sistemas operativos hasta programas de entretención.

El software propietario es otra alternativa, particularmente para el análisis estadístico. Como es bien sabido, este tipo de producto se ofrece, generalmente, en términos de licencia de uso, está protegido por las leyes de Copyright y su desarrollo lo llevan a cabo empresas del rubro. Algunos de estos productos son SPSS, SAS, SYSTAT, Stata, Minitab, XLSTAT, entre muchos otros.

En cualquier escenario, y en el contexto de una investigación de gran envergadura, es muy difícil que un único producto sea capaza de satisfacer todas las necesidades de análisis de datos, y por lo general habrá que apoyar aquellos análisis, así como la comunicación de resultados, en dos o más productos.

Desde el punto de vista del análisis de datos, y en atendiendo a la filosofía que los sustentan, en estas notas se presentan y discuten algunas alternativas libres de productos de software.



Mapa conceptual del software libre. Fuente: wikipedia.com

Muchos de los productos de distribución libre (y de código abierto) se circunscriben al proyecto GNU[1], y se distribuye bajo los términos de licencias GPL (General Public Licence), de las cuales la más conocida es la GNU GPL. Bajo estas licencias, el autor o creador de un producto de software conserva los derechos de autor; pero permite la redistribución del producto, la modificación del código, siempre que el producto resultante siga estando bajo la licencia GNU GPL.

Entre los productos de software libre disponibles en la actualidad para el análisis de datos, destacan el programa R (www.r-project.org), que en los últimos años ha conseguido una importante de cantidad de adeptos, distribuidos entre usuarios personales e institucionales. Este producto es hoy uno de los mejores disponibles. Otro producto de software, distribuido bajo la misma licencia GNU GPL, es el programa PSPP, mismo que describimos en las secciones siguientes.

EL PROGRAMA PSPP

[1] El proyecto GNU fue iniciado por Richard Stallman con el objetivo de crear un sistema operativo completamente libre: el sistema GNU.

El programa PSPP es un software libre, de código abierto. Esto significa que sus potencialidades aumentan con el aporte de usuarios dispuestos a colaborar en su desarrollo, usuarios que forman parte de una extensa comunidad distribuida en todo el mundo, permitiendo y promoviendo el intercambio de información y el trabajo interdisciplinario.

Existen versiones de PSPP para diversas plataformas (Linux, Windows, Mac OS X). En estas notas se utilizó la versión para Windows del programa.

Obtención e instalación del programa

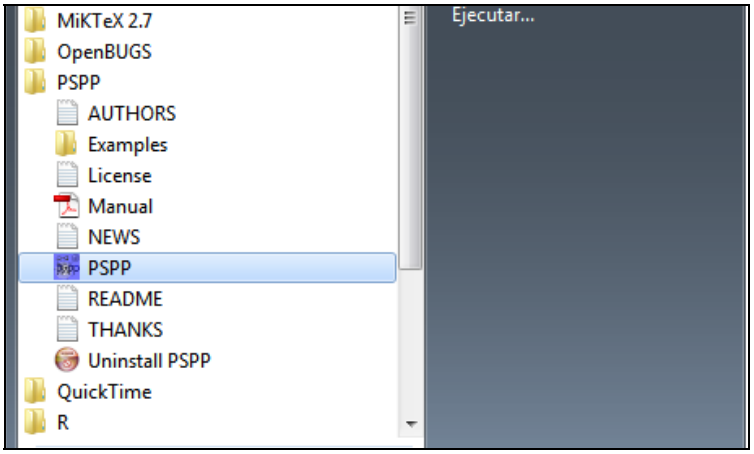
El software se puede descargar desde el sitio www.gnu.org/software/pspp. En este mismo sitio, además, existe diversa documentación, relacionada con plataformas, uso, etc.

La instalación del programa es muy sencilla: simplemente, dando doble click sobre el icono del programa de instalación, se pone en funcionamiento el proceso automático de instalación, proceso que lleva unos breves minutos, tras lo cual se habilita un acceso directo al programa y a un manual en formato pdf (en inglés).

A la fecha de creación de este documento, la versión disponible del software era la 0.8.1, de fecha 01 de noviembre de 2013. En este documento se utilizó la versión 0.8.0, ya que la versión 0.8.1 presenta algunos problemas de formato (no de procesos), especialmente en la barra de iconos.

Puesta en marcha

La puesta en funcionamiento del programa puede hacerse mediante el acceso directo o bien a través del menú Inicio de Windows (figura siguiente).

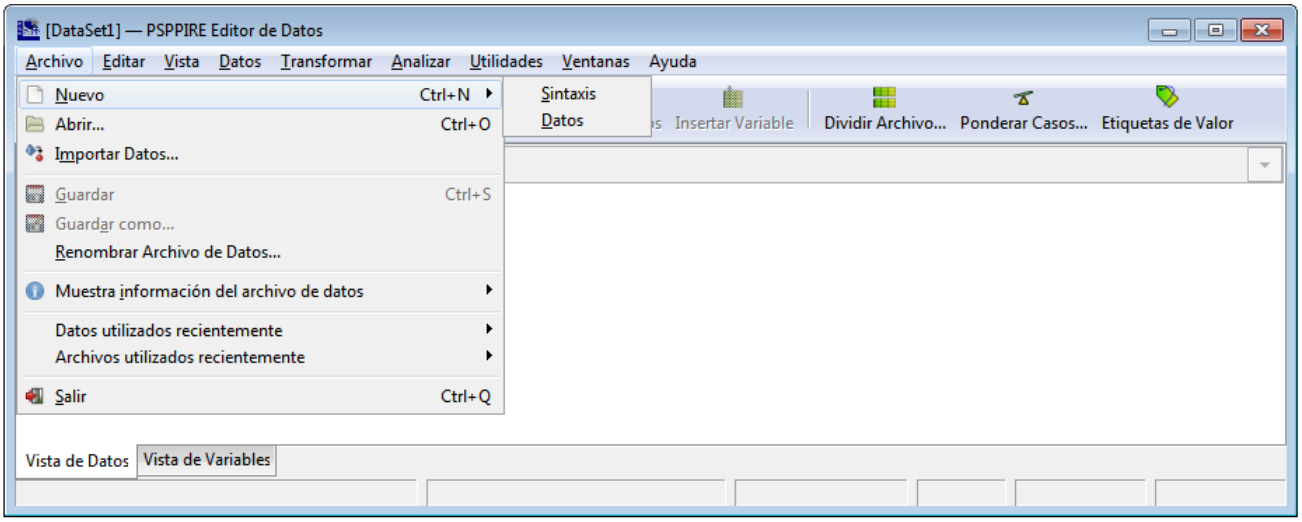


La puesta en funcionamiento de este programa, como ocurre con la mayoría de los software, provee al usuario de una interfaz[2] de comunicación con el programa, que se denomina abreviadamente GUI, por sus siglas en inglés, y que podríamos llamar simplemente "la consola". Éste es un espacio (o ventana) en la que el usuario se comunica con el software. Esta comunicación puede ser, básicamente, de dos formas: mediante una serie de dispositivos (botones, cuadros de diálogo, etc.) pre-programados para llevar a cabo tareas específicas sin más que dar uno o más clicks. La otra forma de comunicación es mediante un diálogo directo, ajustado a la medida del usuario, que usa el lenguaje de programación del software. Es la línea de comandos o la ventana de Sintaxis. Es muy común que los usuarios prefieran la primera forma de interacción con un software, ya que la se-

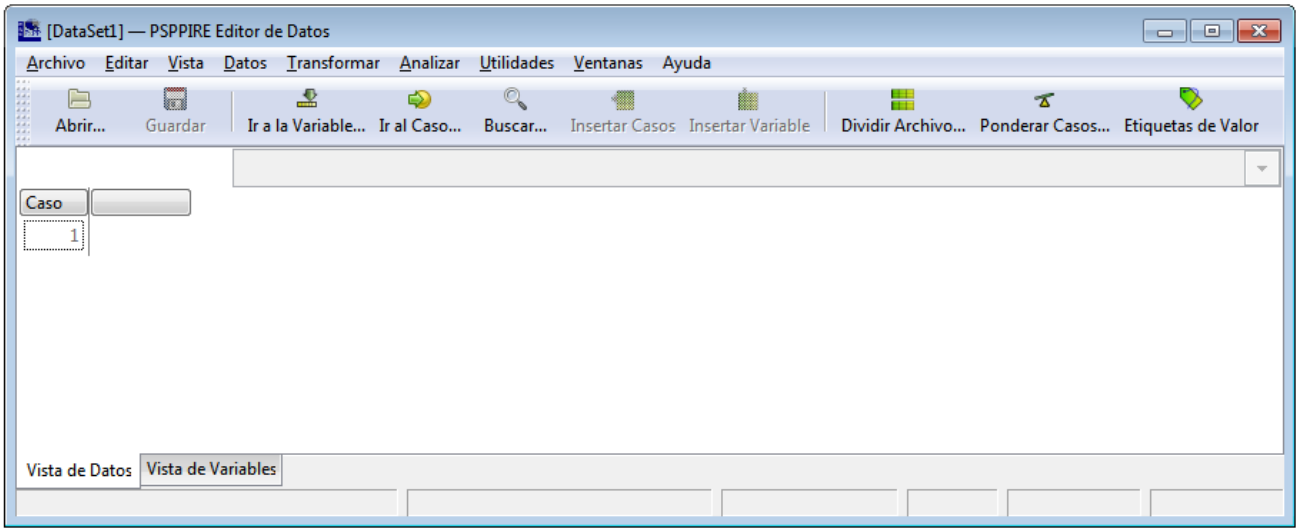
[2] Una interfaz de usuario es cualquier medio de comunicación de un usuario (persona) con un dispositivo, máquina, software, etc.

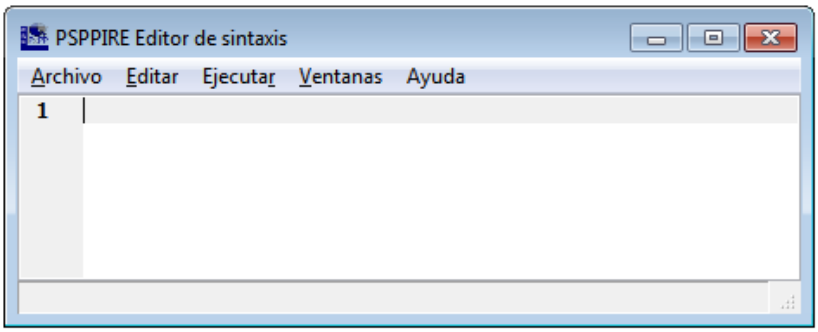
gunda requiere de tener a mano las diversas formas de sintaxis con las que solicitar los análisis respectivos. Sin embargo, el uso y manejo, aunque sea a nivel elemental, de la segunda forma de interacción, le da al usuario un espectro muchísimo más amplio de posibilidades de análisis, y acceso a rutinas o funciones que muchas veces no están disponibles en la parte gráfica de la consola.

El software PSPP provee al usuario de las dos formas antes descritas para interactuar, y las dispone en un mismo dispositivo: la ventana principal del programa, que es también la ventana de Datos. En esta ventana, y particularmente en el menú Archivo, se puede activar la ventana de Datos y la ventana de Sintaxis, tal como se muestra en la figura siguiente, aunque en primer lugar, tal como se comentará más abajo, lo primero que el programa dispone es la ventana de Datos.



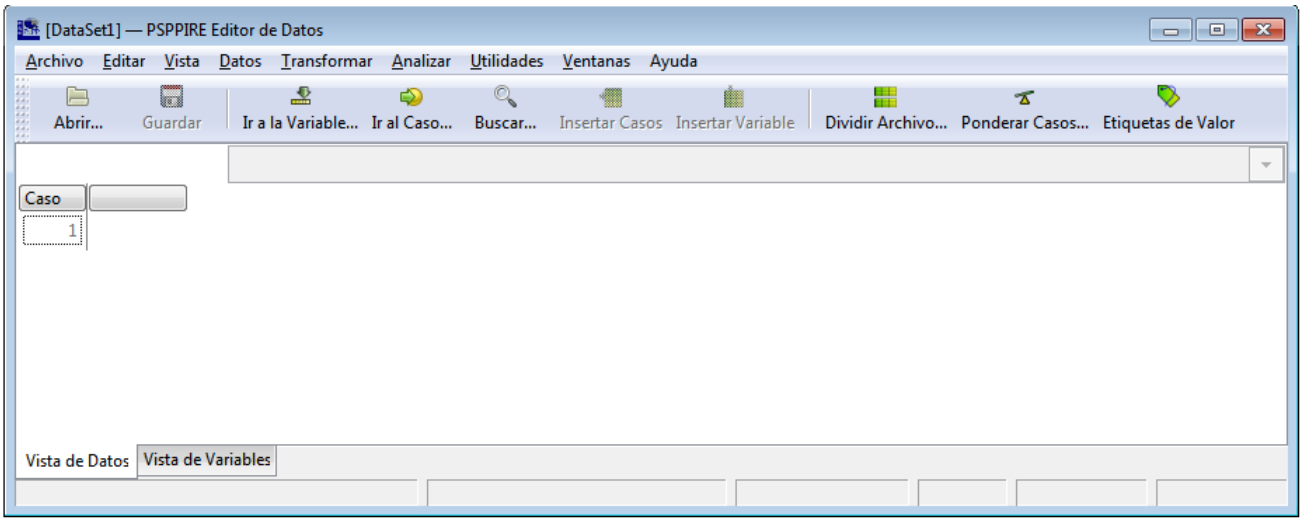
Un aspecto parcial de las ventanas de Datos y de Sintaxis, respectivamente, son las siguientes figuras:





Gestión de datos

Hay básicamente dos situaciones relacionadas con la disponibilidad de datos en PSPP para su posterior análisis: escribir los datos directamente, usando para ello la ventana de Datos, única opción cuando se trata la primera disposición de los datos en formato digital (vaciado de una encuesta a un software). La otra opción, para datos ya digitados en algún otro formato, es la importación de aquel formato externo al formato usado por PSPP. Esta opción es útil cuando, por ejemplo, se tienen datos disponibles en formatos como Excel, ASCII, SPSS, etc. En cualquiera de los dos casos, la disponibilidad de datos se hace en la ventana de Datos, la que tiene una apariencia similar a la de la figura siguiente:



Dejaremos el ingreso directo de datos para otra oportunidad, y ahora nos centraremos en la segunda forma de disponer datos para análisis estadísticos: la importación.

Una de las primeras opciones que se revisan respecto de la gestión de datos de un software es la forma en que éste recupera datos en formatos externos. En la versión aquí utilizada desgraciadamente no funciona la importación desde planillas Excel de Microsoft Office o Calc de LibreOffice. Por esto, y siguiendo los esquemas de importación de datos de gran parte de los programas de software, una de las mejores alternativas es disponer los datos en formato texto, lo que puede hacerse usando un editor de texto plano (ASCII) como el Block de Notas de Windows o alguno similar.

Ilustraremos los diversos pasos del proceso de importación con los datos de HATCO. Estos datos provienen de las encuestas que la "Compañía Hair, Anderson y Tahtam" aplica a sus clientes (se trata de una situación ficticia, y sólo persigue fines pedagógicos). Se encuentra disponible y utilizada en el texto HAIR-ANDERSON-TATHAM-BLACK: ANÁLISIS MULTIVARIANTE. 5° Edición. Prentice-Hall. Madrid, 1999

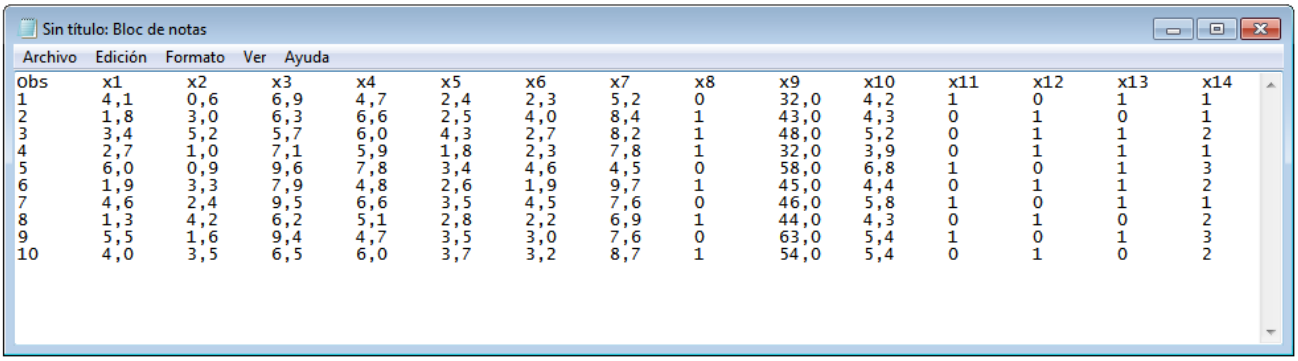
(En las páginas 24 a 26 se describe esta base de datos). Es una base de datos que consiste en 100 observaciones de 14 variables de una empresa llamada HATCO. Se utilizan tres tipos de datos. La primera clase es la percepción de HATCO sobre siete atributos identificados en estudios pasados como los más influyentes en la elección de distribuidor. Los encuestados, ejecutivos de compras de empresas clientes de HATCO, puntúan a HATCO sobre cada atributo. La segunda clase de información hace referencia a los resultados de compras reales, bien sobre las evaluaciones de la satisfacción de los encuestados con HATCO, bien sobre el porcentaje de sus compras de productos a HATCO. La tercera clase de información contiene características generales de las empresas clientes (por ejemplo, tamaño de la empresa, tipo de industria). En resumen, hay disponible en estos datos información cualitativa y cuantitativa. A continuación se proporciona una breve descripción de las variables.

x1	Velocidad de entrega - tiempo que transcurre hasta que se entrega el producto, una vez que se hubo confirmado el pedido.
x2	Nivel de precio - nivel de precios percibido por los clientes industriales.
x3	Flexibilidad de precios - la disposición percibida en los representantes de HATCO para negociar el precio de todas las compras.
x4	Imagen del fabricante - imagen conjunta del fabricante/distribuidor.
x5	Servicio - nivel conjunto de servicio necesario para mantener una relación satisfactoria entre el oferente y el comprador.
x6	Imagen de la fuerza de ventas - imagen conjunta de la fuerza de ventas del fabricante.
x7	Calidad del producto - nivel de calidad percibido en un producto particular (por ejemplo, el acabado o el rendimiento).
x9	Nivel de fidelidad - cuánto se compra a HATCO del total del producto de la empresa, medido en una escala porcentual, que va desde 0 al 100 por cien.
x10	Nivel de satisfacción - satisfacción del comprador con las compras anteriores realizadas a HATCO, medida en la misma escala de clasificación que las variables **x1** a **x7**.
x8	Tamaño de la empresa - tamaño relativo de la empresa respecto a otras empresas en el mismo mercado. Esta variable tiene dos categorías: 0 = pequeña y 1 = grande.
x11	Compra detallada - medida por la cual un comprador particular evalúa cada compra separadamente (análisis del valor total) o en función de una compra detallada, donde se especifican precisamente las características del producto deseado. Esta variable tiene dos categorías: 0 = uso de la compra detallada y 1 = emplea la aproximación al análisis del valor total, evaluando cada compra por separado.
x12	Estructura de la adquisición - método de adquisición/compra de productos a una compañía en particular. Esta variable tiene dos categorías: 0 = adquisición descentralizada y 1 = adquisición centralizada.
x13	Tipo de industria - clasificación de la industria a la que pertenece el comprador del producto. Esta variable tiene dos categorías: 0 = otras industrias y 1 = industria de la clase A.
x14	Tipo de situación de compra - tipo de situación a la que se enfrenta el comprador. Esta variable tiene tres categorías: 1 = nueva tarea, 2 = recompra similar modificada y 3 = recompra simple.

Un aspecto parcial del archivo Calc (Excel) conteniendo los datos es:

Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32,0	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43,0	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48,0	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32,0	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58,0	6,8	1	0	1	3

Como primer paso se copian los datos en formato .txt, lo que puede hacerse “copiando” todos los datos y luego “pegarlos” en un archivo del Block de Notas de Windows. Se obtendrá una estructura como la que sigue:

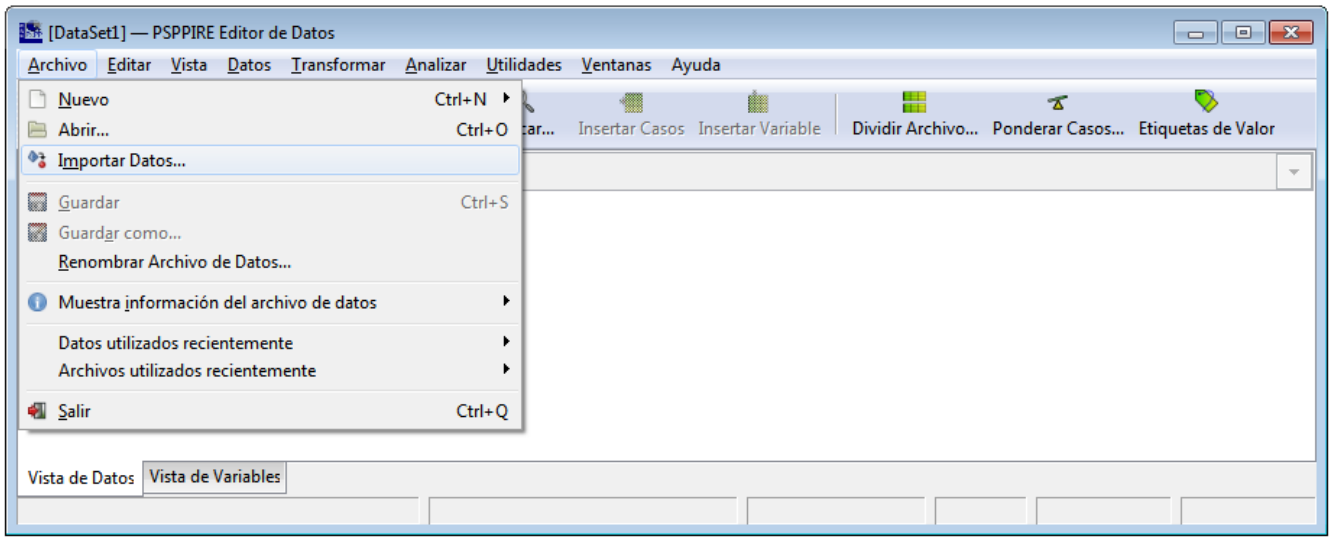


obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32,0	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43,0	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48,0	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32,0	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58,0	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45,0	4,4	0	1	1	2
7	4,6	2,4	9,5	6,6	3,5	4,5	7,6	0	46,0	5,8	1	0	1	1
8	1,3	4,2	6,2	5,1	2,8	2,2	6,9	1	44,0	4,3	0	1	0	2
9	5,5	1,6	9,4	4,7	3,5	3,0	7,6	0	63,0	5,4	1	0	1	3
10	4,0	3,5	6,5	6,0	3,7	3,2	8,7	1	54,0	5,4	0	1	0	2

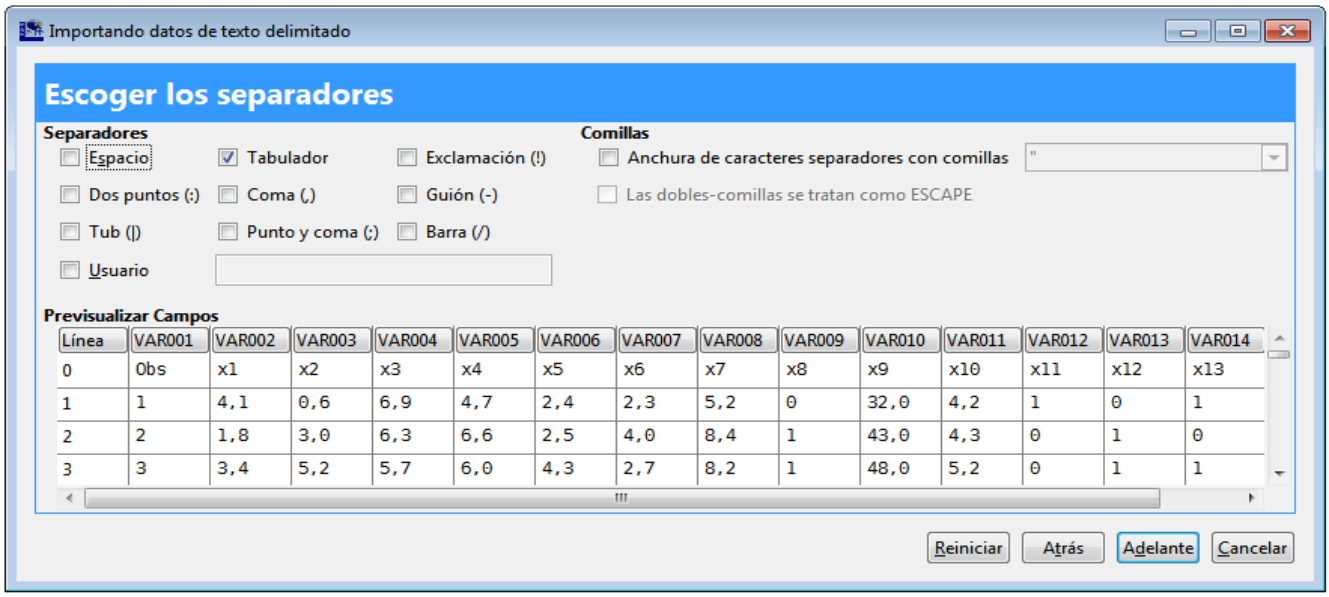
Si la apariencia parece un tanto “desordenada”, no intentar “arreglarla”, porque la estructura de este archivo tiene caracteres de separación y de definición definidos de manera que el software leerá los datos sin problemas.

Seguidamente se graba este archivo de texto.

Luego de lo anterior, y desde el menú Archivo de PSPP, se selecciona la opción Importar datos (figura siguiente).

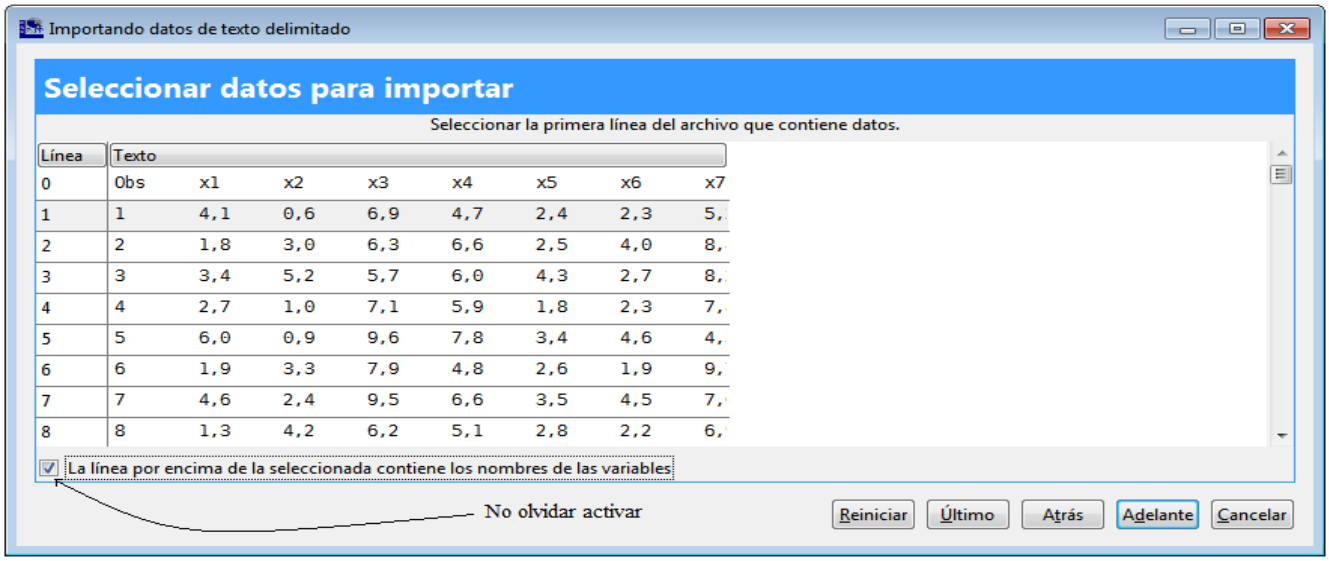


Una vez seleccionada esta opción, se da la ubicación del archivo, y se siguen las indicaciones y procedimientos que van apareciendo. La primera de estas opciones tiene que ver el separador de los campos de los datos.



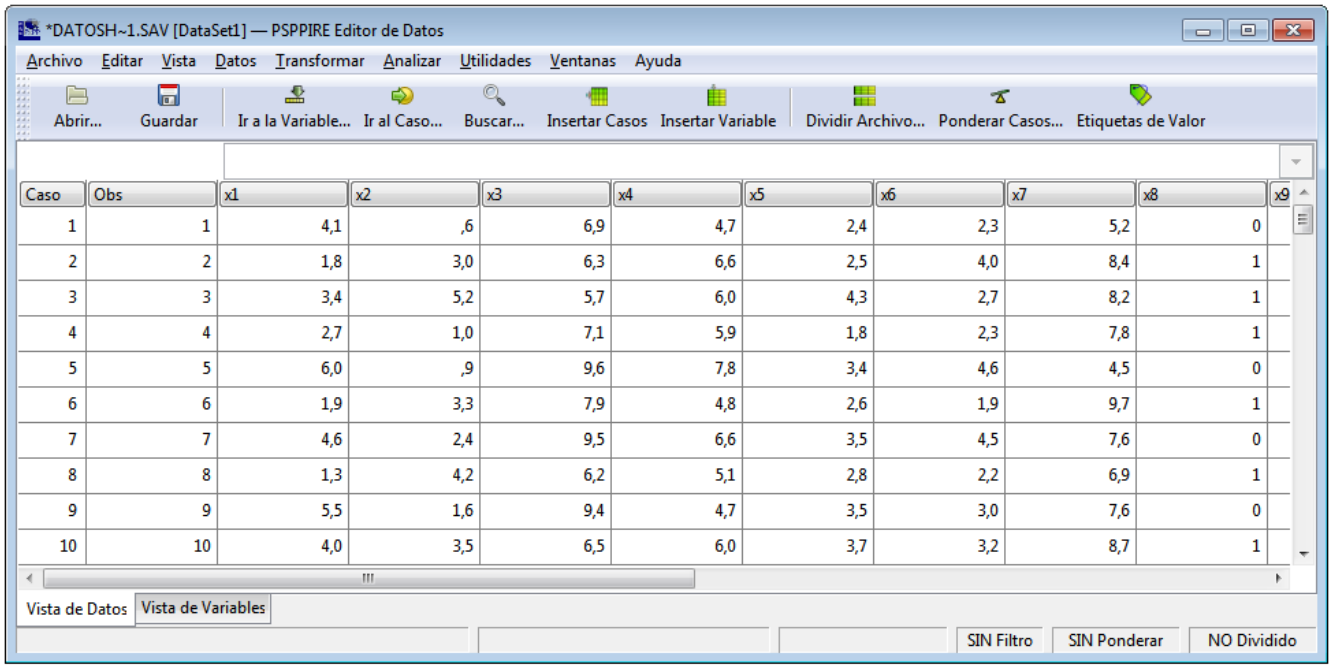
Línea	VAR001	VAR002	VAR003	VAR004	VAR005	VAR006	VAR007	VAR008	VAR009	VAR010	VAR011	VAR012	VAR013	VAR014
0	Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32,0	4,2	1	0	1
2	2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43,0	4,3	0	1	0
3	3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48,0	5,2	0	1	1

Otra de las opciones es `Seleccionar datos para importar` (figura siguiente), que hay que atender con cuidado. En efecto, el programa pide “Seccionar la primera línea del archivo que contiene datos”. Como generalmente la primera línea (fila) de un archivo de datos contiene los nombres de las variables, lo lógico es indicar la segunda línea como la primera línea que contiene los datos, luego de lo cual se debe marcar la casilla de verificación que indica “La línea por encima de la seleccionada contiene los nombres de las variables”, y seguidamente se clicka en el botón Adelante.



Línea	Texto							
0	Obs	x1	x2	x3	x4	x5	x6	x7
1	1	4,1	0,6	6,9	4,7	2,4	2,3	5,
2	2	1,8	3,0	6,3	6,6	2,5	4,0	8,
3	3	3,4	5,2	5,7	6,0	4,3	2,7	8,
4	4	2,7	1,0	7,1	5,9	1,8	2,3	7,
5	5	6,0	0,9	9,6	7,8	3,4	4,6	4,
6	6	1,9	3,3	7,9	4,8	2,6	1,9	9,
7	7	4,6	2,4	9,5	6,6	3,5	4,5	7,
8	8	1,3	4,2	6,2	5,1	2,8	2,2	6,

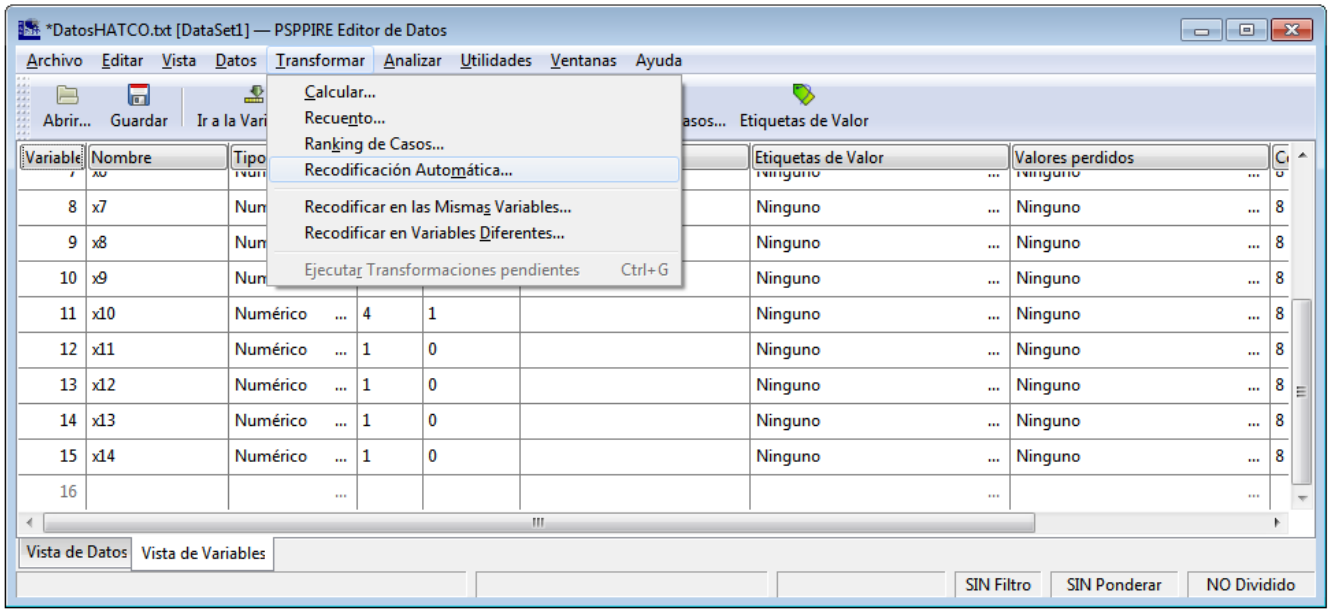
Al clikcar en Aplicar finaliza el proceso de importación, con un resultado similar al que se muestra a continuación:



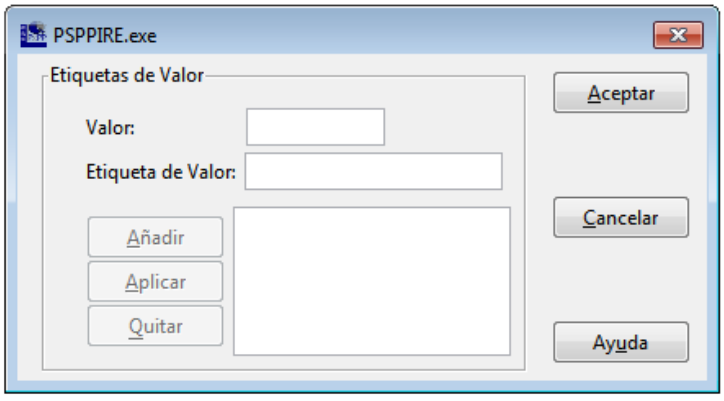
Caso	Obs	x1	x2	x3	x4	x5	x6	x7	x8	x9
1	1	4,1	,6	6,9	4,7	2,4	2,3	5,2	0	
2	2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	
3	3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	
4	4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	
5	5	6,0	,9	9,6	7,8	3,4	4,6	4,5	0	
6	6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	
7	7	4,6	2,4	9,5	6,6	3,5	4,5	7,6	0	
8	8	1,3	4,2	6,2	5,1	2,8	2,2	6,9	1	
9	9	5,5	1,6	9,4	4,7	3,5	3,0	7,6	0	
10	10	4,0	3,5	6,5	6,0	3,7	3,2	8,7	1	

Es muy común que un archivo de datos contenga tanto información cuantitativa como cualitativa. Pero los datos cualitativos casi siempre necesitan codificarse adecuadamente de modo de facilitar y hacer pertinentes los análisis a la naturaleza de los datos. En este sentido es común que estos datos estén expresados en códigos numéricos, como 1 para sexo femenino y 2 para sexo masculino, por ejemplo, en cuyo caso hay que indicar al programa qué significan estos códigos 1 y 2. También puede ocurrir que los datos estén expresados en formato “string” (o cadena de caracteres), como “Femenino” y “Masculino” para referirse al sexo de las unidades de observación. En este caso hará falta también una recodificación de estos valores textuales en valores en formato adecuado para ser tratados estadísticamente. Muchas veces el uso de códigos numéricos es una alternativa satisfactoria.

La etiquetación (codificación) de una variable cualitativa en PSPP se hace en la pestaña Vista de variables del software, en la parte inferior de la ventana principal del programa. Por otro lado, la conversión de una cadena de caracteres en valores de una variable cualitativa puede mediante las opciones del menú Transformar, comando Recodificación Automática… (figura siguiente).



En el archivo que hemos importado, las variables x8, x11, x12, x13 y x14 son todas cualitativas, con códigos 0 (Pequeña) y 1 (Grande) para x8 (tamaño de la empresa), y así en las otras. La codificación de esta variable se hace en la Vista de variables, parte inferior izquierda de la ventana de Datos, completando adecuada y convenientemente los campos correspondientes del cuadro de diálogo que se desplegará, y particularmente el campo Etiquetas de Valor, tal como en la figura siguiente. En el campo Valor se anota el código numérico de la variable, por ejemplo 0, y en el de Etiqueta de Valor la cadena de caracteres correspondiente, en el ejemplo, Pequeña. Seguidamente se clicka en Añadir y se continúa con el siguiente valor de la variable. Cuando se haya completado la codificación, se clicka en Aceptar.

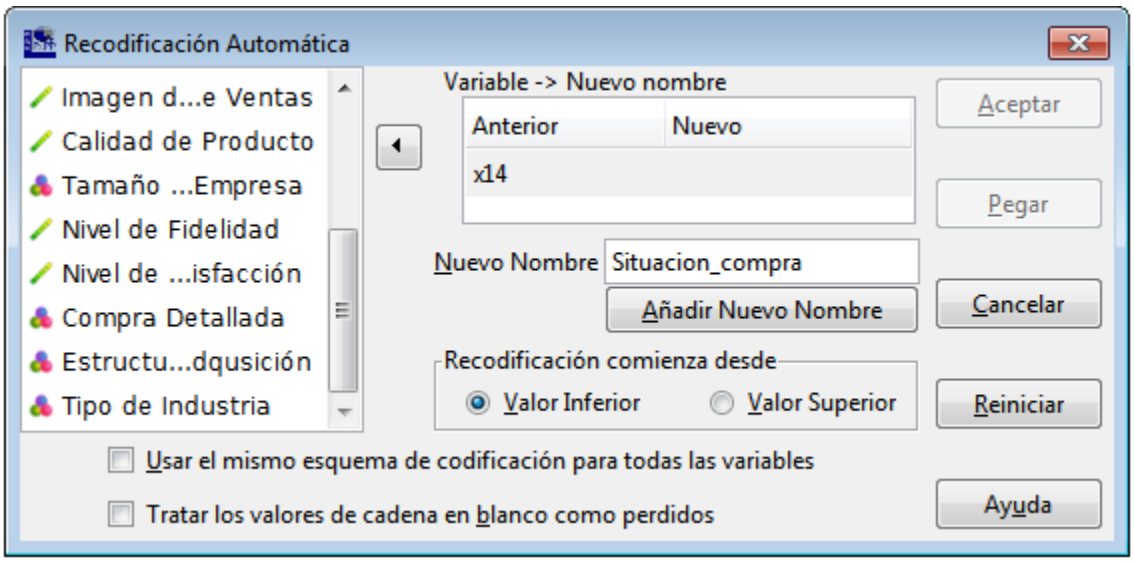


Cuidar de cambiar la Medida de la variable (a Nominal).

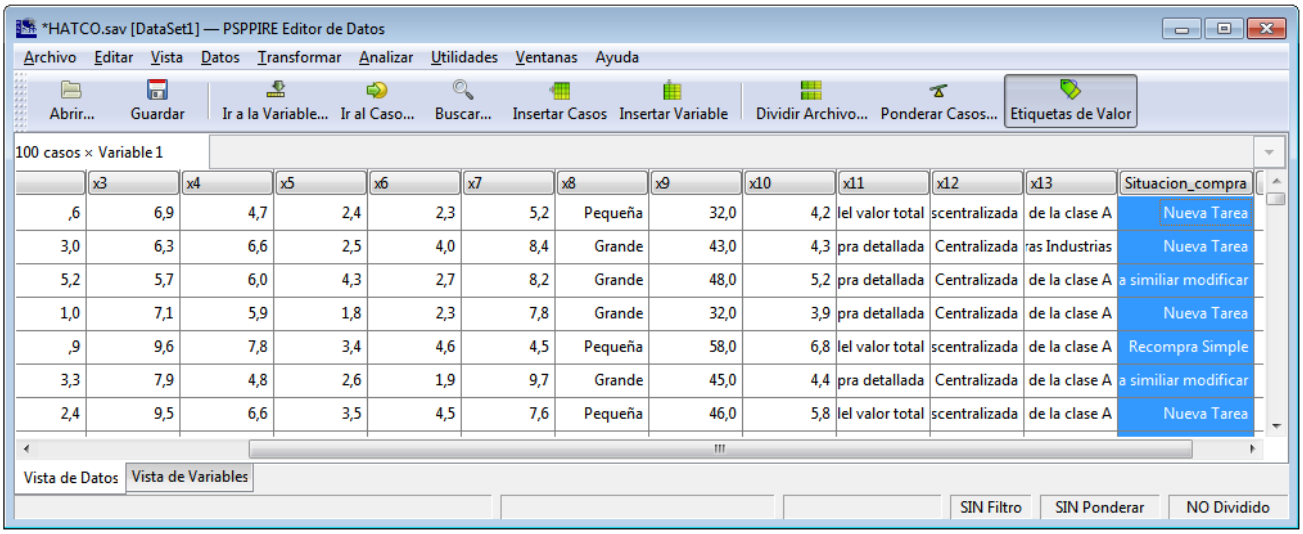
La otra situación se da cuando se tiene una variable cuyos valores están expresados en caracteres (cadena). En esta situación se desea recodificar esta variable de modo que la cadena asociada a cada valor se corresponda con un código numérico. Para ilustrar esta forma considérese la variable x14, y supongamos que los valores con los que aparece en el archivo de datos original son:

nueva_tarea; recompra_similar_modificada; recompra_simple.

La codificación de variables como esta se facilita considerablemente con la opción Recodificación Automática. La ventana de diálogo de esta opción es como la siguiente:

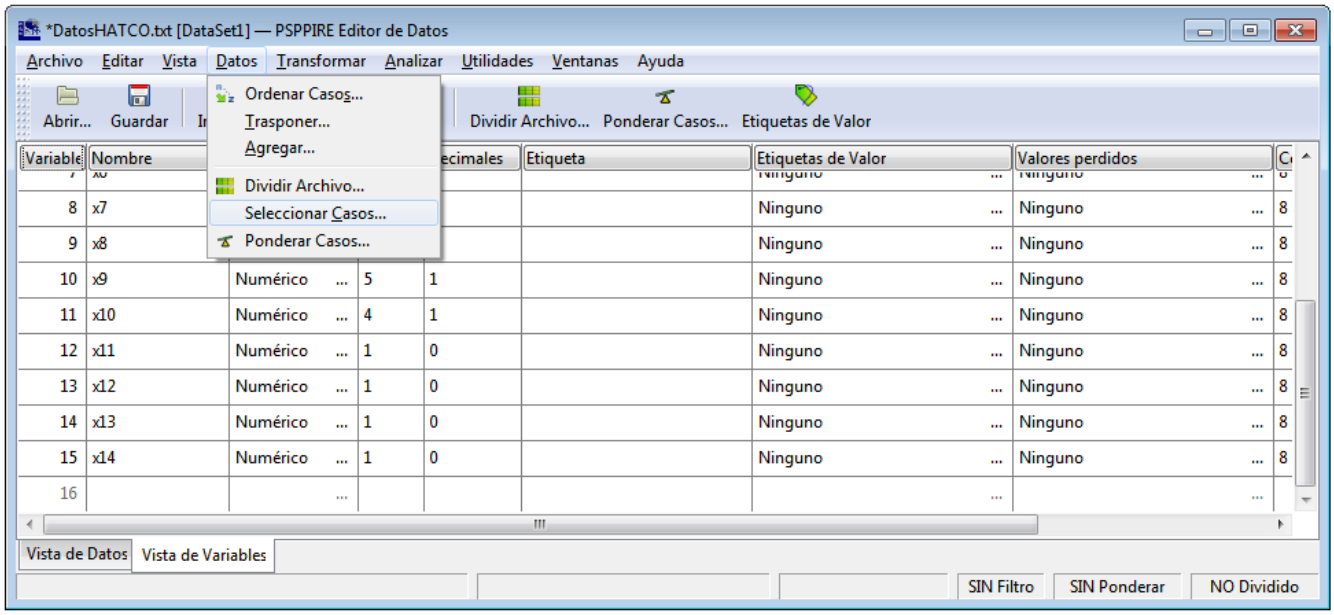


En el campo de la izquierda se busca la variable a recodificar (aquí se ha seleccionado x8, tamaño de la empresa) y se "pasa" al campo de Variable -> Nuevo nombre de la derecha. Seguidamente se completa el campo Nuevo Nombre con un nombre para esta nueva variable (que contendrá la codificación) y se clicka el botón Añadir Nuevo Nombre, luego de lo cual se clicka el botón Aceptar. La acción agrega, al final del archivo, una nueva variable (columna) de nombre Situacion_compra, como muestra la figura siguiente.

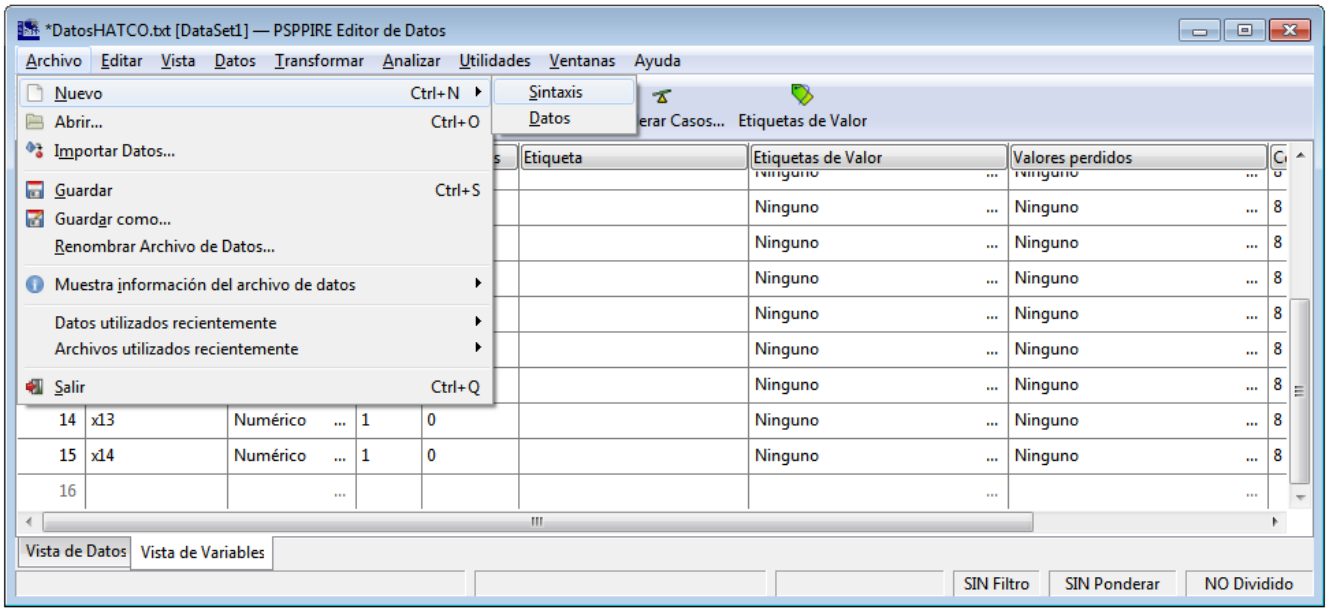


Selección de casos de un archivo de datos

En muchas situaciones se requiere de restringir los análisis a un determinado subconjunto de datos (segmentar el archivo de datos) que satisfacen determinada condición, como por ejemplo que los casos a considerar sean personas de sexo femenino. Algunas de estas posibilidades están disponibles en la opción Seleccionar casos del menú Datos (figura siguiente).

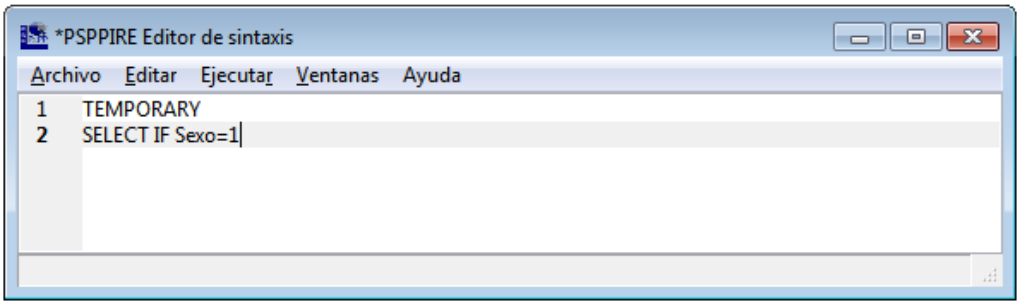


Sin embargo, aún no está disponible esta opción en la interfaz gráfica del programa (ventana de Datos), aunque sí lo está vía la ventana de Sintaxis. Para ilustrar el uso de esta forma de seleccionar un conjunto de casos que satisfagan cierta condición, supóngase que la variable Sexo contiene los códigos 1 para Femenino, y 2 para Masculino, y que deseamos seleccionar todos aquellos casos de Sexo Femenino. Para hacer esto se abre una ventana de sintaxis, lo que se hace en el menú Archivo, opción Nuevo y seguidamente Sintaxis (figura siguiente)



Esta secuencia hace disponible en la ventana Editor de sintaxis. En esta ventana se escriben las siguientes líneas, tal como se muestra en la figura que sigue:

```
TEMPORARY
SELECT IF Sexo=1
```



Luego, en el menú `Ejecutar` de esta ventana, se clicka en Todos.

En la opción Seleccionar casos del menú Datos hay otras formas de seleccionar casos que pudieran ser de interés en determinadas circunstancias.

Análisis de datos

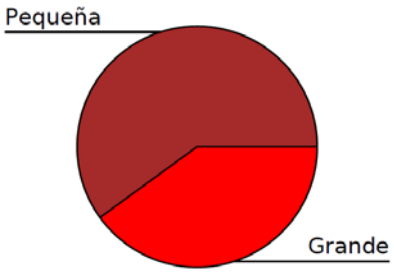
PSPP es un software en desarrollo, lo cual significa que hay muchas funciones básicas que no están aún disponibles vía interfaz gráfica, y algunas que, estando disponibles, aún no tienen el nivel de desarrollo que debieran tener. Por ejemplo, no hay todavía un menú para construir gráficos de barras, ni simples ni agrupadas, gráficos estos útiles para la representación de datos categóricos. En el menú Analizar-Explorar se pueden obtener gráficos circulares. No obstante desde la ventana de sintaxis sí es posible obtener una gama importante de opciones que no están disponibles desde la ventana principal del software.

En el menú Analizar se encuentra la opción Estadística Descriptiva, mediante la cual se puede desarrollar estadística exploratoria y descriptiva, con algunas opciones de gráficos en ciertos casos. A la fecha de elaboración de este documento aún no existían más opciones en la interfaz gráfica para la elaboración de gráficas asociadas al análisis de datos; pero el tratamiento a nivel numérico parece bastante bueno y completo. No obstante, desde la ventana de Sintaxis se pueden obtener algunas representaciones, como histogramas y boxplot, además de estadísticas descriptivas e inferenciales.

Un ejemplo

Nada mejor que un ejemplo para ilustrar la puesta en práctica de un software. Para ello se considera el archivo de datos HATCO, y nos proponemos explorar el Nivel de satisfacción de los clientes (x10) en función del tamaño de la empresa cliente (x8). Informar de esta situación implica dos procesos: Explorar el nivel de satisfacción y establecer si las diferencias observadas son o no significativas. Lo primero es Estadística Descriptiva, mientras que lo segundo, Estadística Inferencial. Para llevar a cabo cualquiera de estos procesos o análisis, lo primero a realizar es disponer de un archivo de datos en formato interpretable por PSPP, es decir, importar los datos a PSPP desde su formato original y hacer las codificaciones correspondientes, lo que ya se hizo según lo explicado más arriba. Por lo tanto, suponemos que los datos están disponibles en un archivo PSPP, y que hemos nombrado como “HATCO.sav” (la extensión .sav es propia del programa SPSS, programa propietario muy conocido y de amplio uso en el mundo académico. Esto le da al PSPP una de las más importantes características: compatibilidad casi total con SPSS).

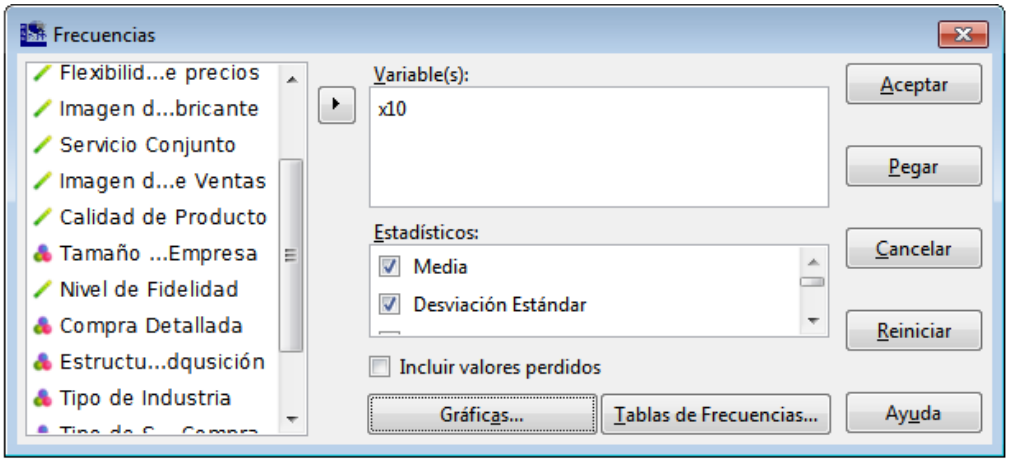
Cuando se quiere hacer un informe de determinada información, como el ejemplo que estamos analizando, una de las primeras partes de dicho informe debe mostrar la información disponible. En este caso debería mostrarse la distribución de las empresas estudiadas según su tamaño (composición de la muestra), así como la característica observada. En relación a la conformación de la muestra, digamos que en la muestra hay 60 empresas pequeñas y 40 empresas grandes, distribución resumida en el gráfico adjunto (y obtenida siguiendo la ruta `Analizar->Estadística Descriptiva->Frecuencias`…)



Seguidamente mostraremos, también mediante recursos gráficos, la distribución global de la variable Nivel de satisfacción (x10). Esto lo podemos hacer siguiendo la secuencia

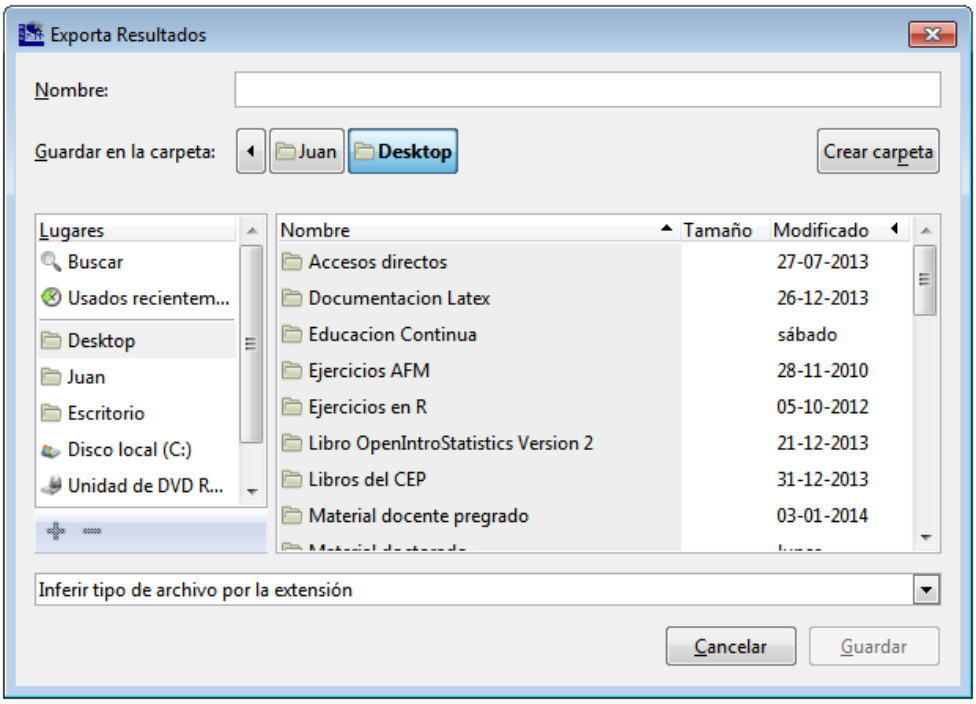
`Analizar->Estadística Descriptiva->Frecuencias`…

acción que despliega un cuadro de diálogo similar al siguiente.



En este cuadro seleccionamos la variable x10 y la “pasamos” al campo Variable(s): de la derecha. En las opciones de Gráficas… marcamos la casilla Dibujar histogramas. Si se desea también, aquí mismo, puede marcarse la casilla Sobreimprimir curva normal. En las opciones del botón Tablas de Frecuencias…, sección Mostrar tabla de frecuencias, elegimos Nunca, puesto que, por tratarse de una variable continua, el sistema construiría una tabla en la que registra la frecuencia de cada observación. Precisamente ésta es la razón por la que se está haciendo un histograma. Finalizada la completación de los distintos campos, se clicka en Aceptar, con lo cual los resultados se disponen en la Ventana de resultados de PSPP.

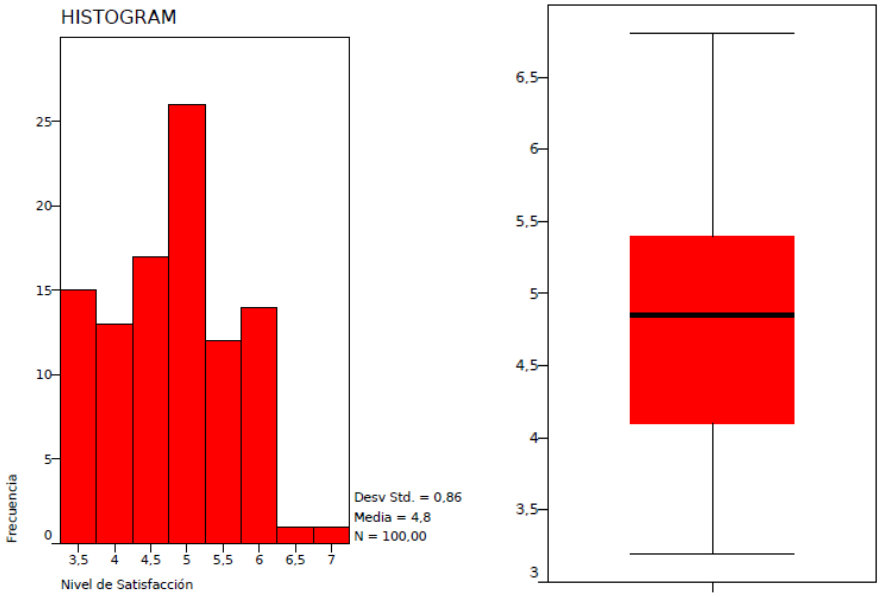
Si no se cierra, la ventana de Resultados acumula la información proveyendo de los distintos análisis, la que se puede guardar usando la opción Exportar del menú Archivo de esta ventana, que requiere de indicar nombre, ubicación y tipo de archivo (figura siguiente). Quizá sea necesario utilizar recursos adicionales para obtener partes específicas de este archivo de resultados. En estas notas, por ejemplo, se optó por guardar en formato pdf y aquí utilizar herramientas de edición para “recortar” las zonas del archivo que nos interesaban.



Si lo deseado es mostrar la distribución global de x10 mediante un diagrama de cajas (Box-plot), entonces hay que actuar en la ventana de Sintaxis, con la siguiente sintaxis:

```
EXAMINE
/VARIABLES = x10
/PLOT = BOXPLOT
```

Ambos resultados se muestran a continuación.



Ahora exploraremos la misma variable x10, pero en función de la variable x8 (Tamaño de la empresa). La secuencia

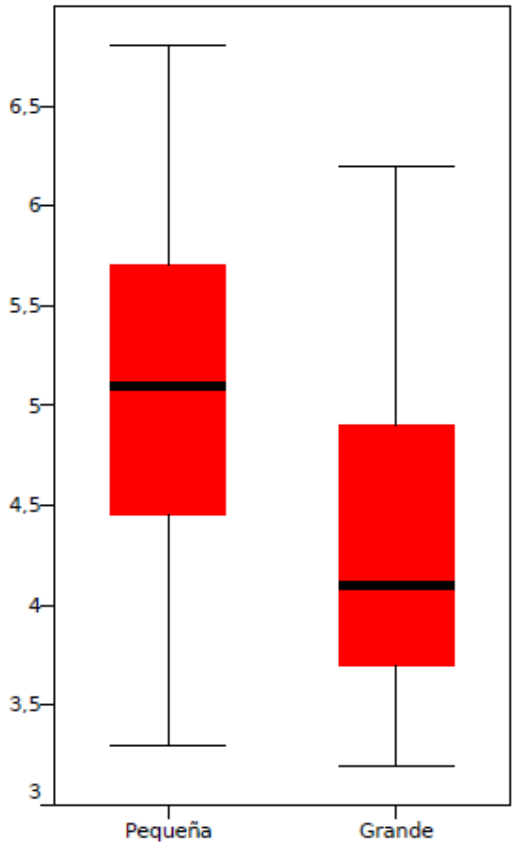
`Analizar->Estadística Descriptiva->Explorar…`

entrega solamente resultados numéricos.

Para mostrar gráficamente la comparación, mediante un box-plot, hay que utilizar la ventana de Sintaxis, con las siguientes líneas:

```
EXAMINE
/VARIABLES = x10 BY x8
/PLOT = BOXPLOT
```

El resultado es la figura adjunta.



Estos últimos análisis, junto a la información numérica provista por el software, permiten adelantar (por no decir 'concluir') que existe importante evidencia para concluir que, en general, el nivel de satisfacción es mayor en las empresas pequeñas. Aquí, además, la distribución del nivel de satisfacción evidencia un importante grado de simetría, lo que le da validez al uso del promedio y la desviación estándar como medidas de resumen de los datos. Por su parte, en cambio, el nivel de satisfacción de las empresas grandes, aparte de ser menor que en de las empresas pequeñas, presenta un alto grado de asimetría, prevaleciendo una valoración más bien baja del grado de satisfacción.

Sin embargo, el aparente grado tanto de simetría en empresas pequeñas, como de asimetría en el caso de empresas grandes, debe ser medido estadísticamente, lo cual es parte del aspecto inferencial del análisis, que pasamos a estudiar a continuación.

En un contexto inferencial, lo que esperamos es establecer si el nivel de satisfacción de los clientes es diferente según el tamaño de la empresa. Pero esto corresponde a contrastar las hipótesis:

$$H_0 : \mu_P = \mu_G \text{ v/s } H_a : \mu_P = \mu_G$$

donde μ_P es el nivel de satisfacción promedio, a nivel poblacional, de las empresas pequeñas; y μ_G es el nivel de satisfacción promedio, a nivel poblacional, de las empresas grandes.

Como es sabido, este contraste requiere de normalidad en cada conjunto de datos (en las empresas pequeñas y en las empresas grandes). Para el caso de una sola muestra, PSPP provee de pruebas de normalidad en

`Analizar->Pruebas No-paramétricas->K-S para Una Muestra…`

Desgraciadamente en nuestro ejercicio tenemos dos grupos o muestras.

Una forma de llevar a cabo ambas pruebas es seleccionar primero las empresas pequeñas, y luego las grandes, y para cada caso llevar a cabo sendas pruebas de normalidad. Esto puede hacerse en la ventana de Sintaxis, mediante la ejecución de las siguientes líneas de comandos:

TEMPORARY
SELECT IF x8= (indicar 0 o bien 1)
NPAR TESTS
/K-S (NORMAL) = x10

Los resultados son:

```
Prueba Kolmogorov_Smirnov para una muestra
#=========================================#====================#
#                                         |Nivel de Satisfacción#
#=========================================#====================#
#N                                        |                  60#
#Parámetros Normal        Media           |                5,09#
#                         Desviación Estándar|                ,75#
#Diferencias Más Extremas Absoluto        |                ,08#
#                         Positivo        |                ,07#
#                         Negativo        |               -,08#
#Z de Kolmogorov-Smirnov                  |                ,60#
#Sig. Asint. (2-colas)                    |                ,87#
#=========================================#====================#

Prueba Kolmogorov_Smirnov para una muestra
#=========================================#====================#
#                                         |Nivel de Satisfacción#
#=========================================#====================#
#N                                        |                  40#
#Parámetros Normal        Media           |                4,29#
#                         Desviación Estándar|                ,78#
#Diferencias Más Extremas Absoluto        |                ,13#
#                         Positivo        |                ,13#
#                         Negativo        |               -,08#
#Z de Kolmogorov-Smirnov                  |                ,80#
#Sig. Asint. (2-colas)                    |                ,54#
#=========================================#====================#
```

Observando el valor-p en ambos análisis, puede concluirse que en ambos tipos de empresa el nivel de satisfacción se distribuye normalmente. Por lo tanto dicho nivel es comparable en esto grupos de empresas. Esta comparación debe hacerse mediante una prueba T, que será T-Student si las varianzas de ambos grupos son homogéneas para el nivel de satisfacción, o bien será T-Welch cuando las varianzas sean heterogéneas. Esta información la provee PSPP como parte de los resultados del contraste.

El contraste de las hipótesis que estamos estudiando se lleva a cabo siguiendo la siguiente secuencia:

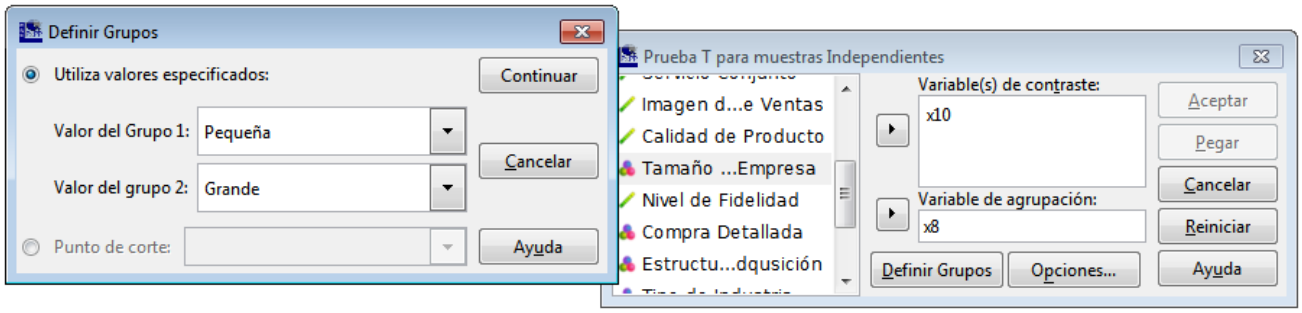
`Analizar->Comparar Medias->Prueba T para Muestras Independientes…`

En los cuadros y ventanas sucesivas que aparecen, se completan los campos con la información requerida (ver figura siguiente). Los resultados más relevantes del proceso se resumen en la tabla que sigue.

	Prueba Levene		Prueba T para la igualdad de medias		
	F	Sig.	t	df	Sig.
Se asume homogeneidad	0,32	0,57	5,08	98	0,00
No se asume homogeneidad			5,04	81,64	0,00

Estos resultados evidencian, en primer lugar, que las varianzas de ambos grupos son homogéneas, debiendo usar la T-Student para comparar las medias. Esta última comparación conduce, según estos datos, al rechazo de la hipótesis nula (valor-p igual a 0,00). En consecuencia debe concluirse que el nivel de satisfacción de las empresas pequeñas es significativamente diferente al de las empresas grandes. Los promedios observados, así como las representaciones gráficas iniciales, permiten concluir, con un 95% de confianza (nivel que usa PSPP

en sus análisis) que el nivel de satisfacción de la empresas grandes es menor que el de las empresas pequeñas. La empresa HATCO quizá deba pensar una estrategia orientada a las empresas grandes con tal de aumentar su nivel de satisfacción.



Resumen de principales opciones de análisis de PSPP

En la tabla que sigue se resumen las principales funciones de PSPP, no disponibles en la GUI, para análisis de datos. En paréntesis tipo corchete, opcional; en paréntesis de llave de conjunto, alternativas posibles. STATISTICS puede ser MEAN, SEMEAN, MEDIAN, STDDEV, VARIANCE, SUM, MINIMUN, MAXIMUN.

PROCEDIMIENTO/TAREA	SINTAXIS
Gráfico de cajas	EXAMINE /VARIABLES = Variable a explorar [BY Variable factor] /PLOT = BOXPLOT
Gráfico de normalidad	EXAMINE /VARIABLES = Variable a explorar [BY Variable factor] /PLOT = NPPLOT
Histogramas (para comparar)	EXAMINE /VARIABLES = Variable a explorar BY Variable factor /PLOT = HISTOGRAM
Análisis de la Varianza One-Way	ONEWAY /VARIABLES = Lista de variables respuesta BY Factor /STATISTICS = {DESCRIPTIVES, HOMOGENEITY} /POSTHOC = {LSD, SCHEFFE, TUKEY}

Temuco, enero de 2014