CURSO SPSS

BLOQUE I – DATOS Y CASOS

BLOQUE II – ANÁLISIS DE DATOS

TEMA - 1

Entrada de datos y definición de variables

1.- Entrada de datos

La ventana de introducción de datos nos permite introducir los datos que pretendemos analizar, o visualizar los datos previamente introducidos. De igual modo en ella se encuentran las utilidades que nos permiten importar los datos desde otros formatos como pueden ser datos ASCII, una hoja de cálculo de EXCEL, etc.

1.1.- Lectura de archivos de Excel.

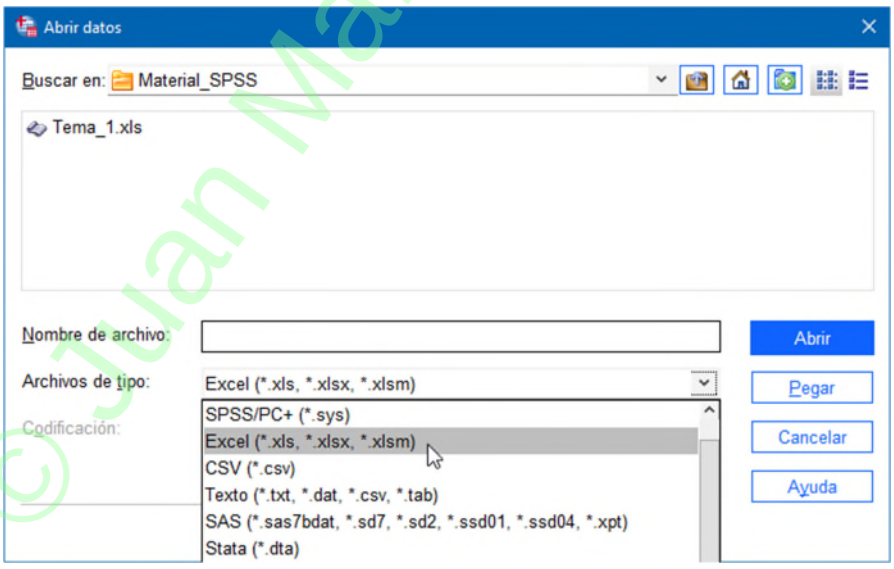
En primer lugar, hay que tener en cuenta las siguientes observaciones:

Tipo y ancho de datos. Cada columna es una variable. El tipo de datos y el ancho de cada variable están determinados por el tipo de datos y el ancho en el archivo de Excel. Si la columna contiene más de un tipo de datos (por ejemplo, fecha y numérico), el tipo de datos se define como cadena y todos los valores se leen como valores de cadena válidos.

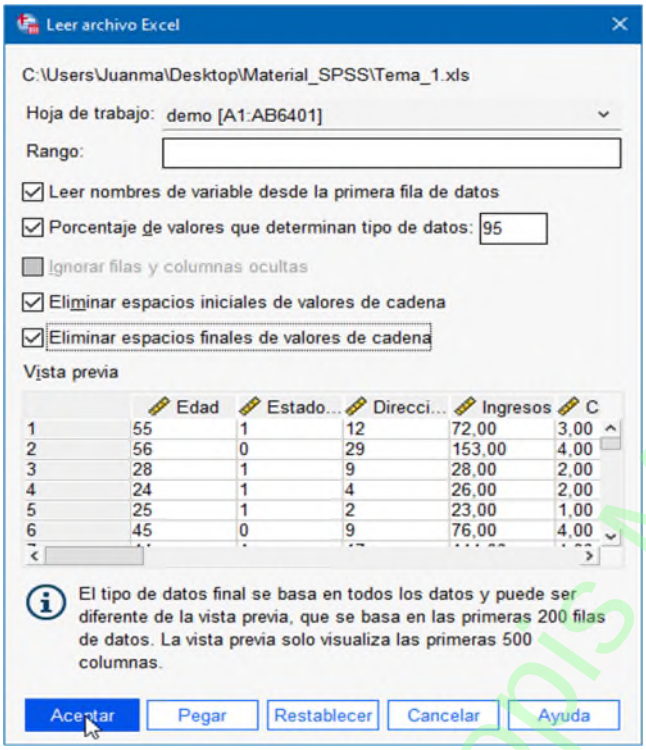
Casillas en blanco. En las variables numéricas, las casillas en blanco se convierten en el valor perdido del sistema indicado por un punto (o una coma). En las variables de cadena, los espacios en blanco son valores de cadena válidos y las casillas en blanco se tratan como valores de cadena válidos.

El procedimiento para leer datos de una hoja de un archivo Excel, es el siguiente:

Seleccionar en el menú superior: Archivo --- Abrir --- Datos, aparecerá la siguiente pantalla:



Elegir el tipo "Excel (*.xls)" de manera que aparecerán los archivos con esa extensión (si los hay). A continuación, seleccionar y abrir el archivo deseado. Aparecerá un cuadro de diálogo como el presentado a continuación:

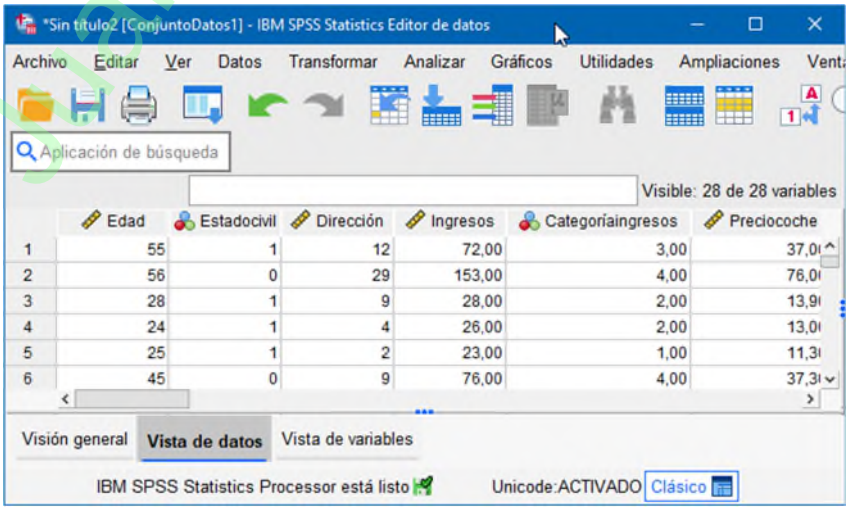


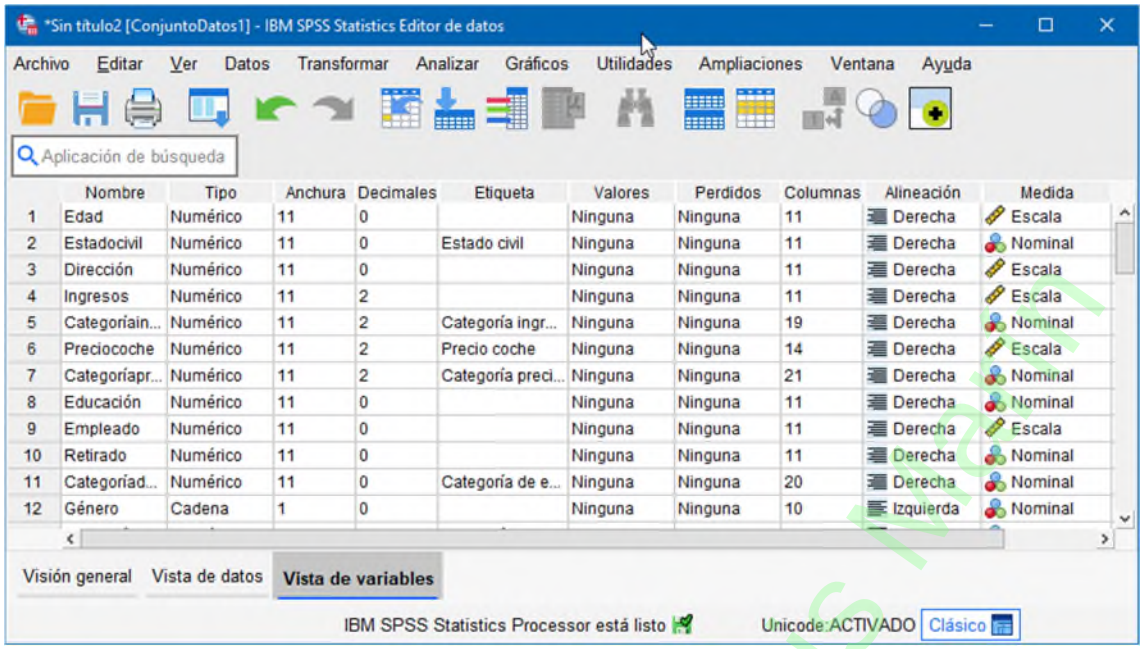
Leer nombres de variables de la primera fila de datos. Si lee la primera fila del archivo de Excel (o la primera fila del rango especificado) como nombres de variable, los valores que no cumplan las normas de denominación de variables se convertirán en nombres de variables válidos y los nombres originales se utilizarán como etiquetas de variable. Si no lee nombres de variable del archivo de Excel, se asignarán nombres de variable por defecto.

Hoja de trabajo. Los archivos de Excel pueden contener varias hojas de trabajo. El Editor de datos lee por defecto la primera hoja. Para leer una diferente, seleccione la que desee en la lista desplegable.

Rango. También puede leer un rango de casillas. Para especificar rangos de casillas utilice el mismo método que emplearía en Excel.

Al pulsar el botón *Aceptar*, los datos aparecerán en la ventana de introducción de datos.





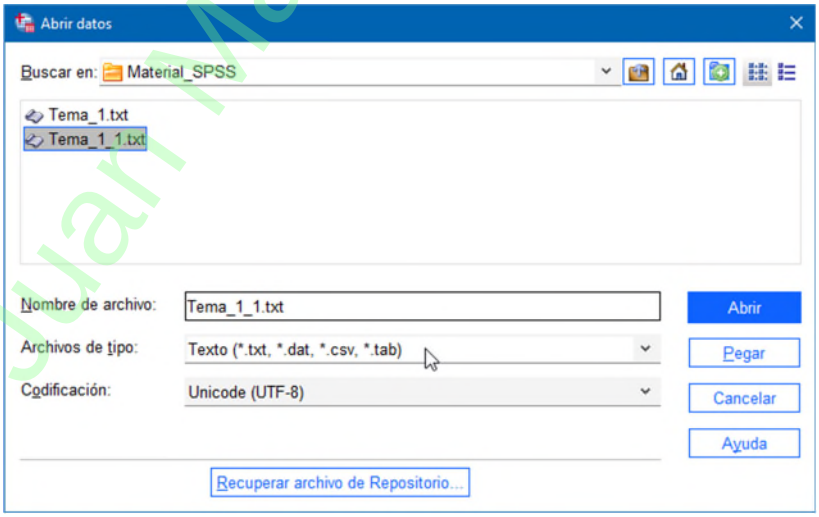
1.2.- Lectura de archivos de texto

El Asistente para la importación de texto puede leer archivos de datos de texto de diversos formatos:

- Archivos delimitados por tabuladores
- Archivos delimitados por espacios
- Archivos delimitados por comas
- Archivos con formato de campos fijos

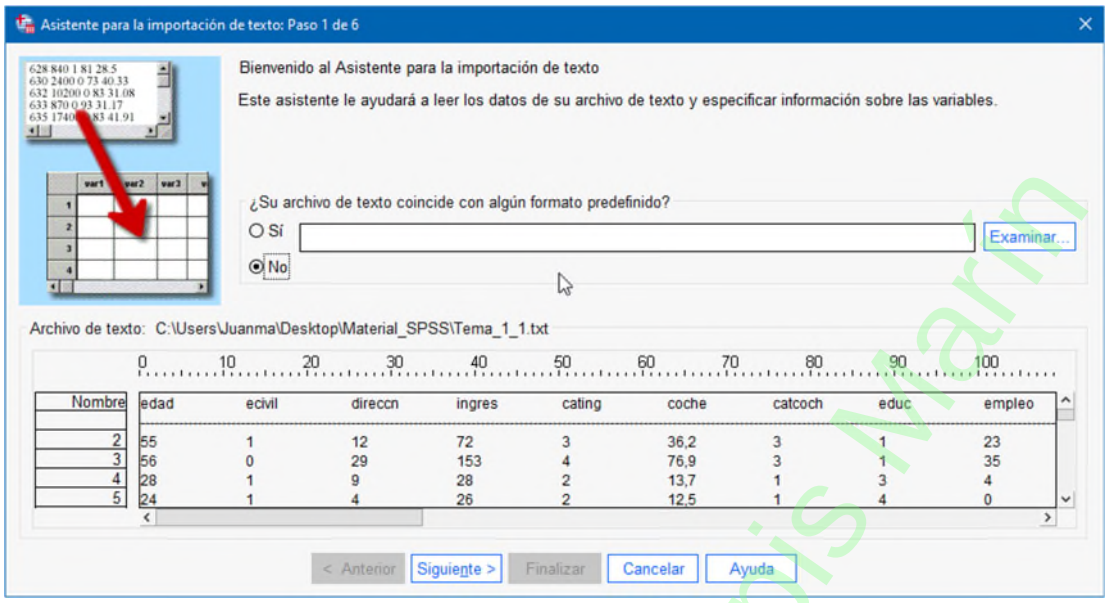
Para leer archivos de datos de texto

- Elija en los menús: Archivo --- Abrir --- Datos, aparecerá la siguiente pantalla:

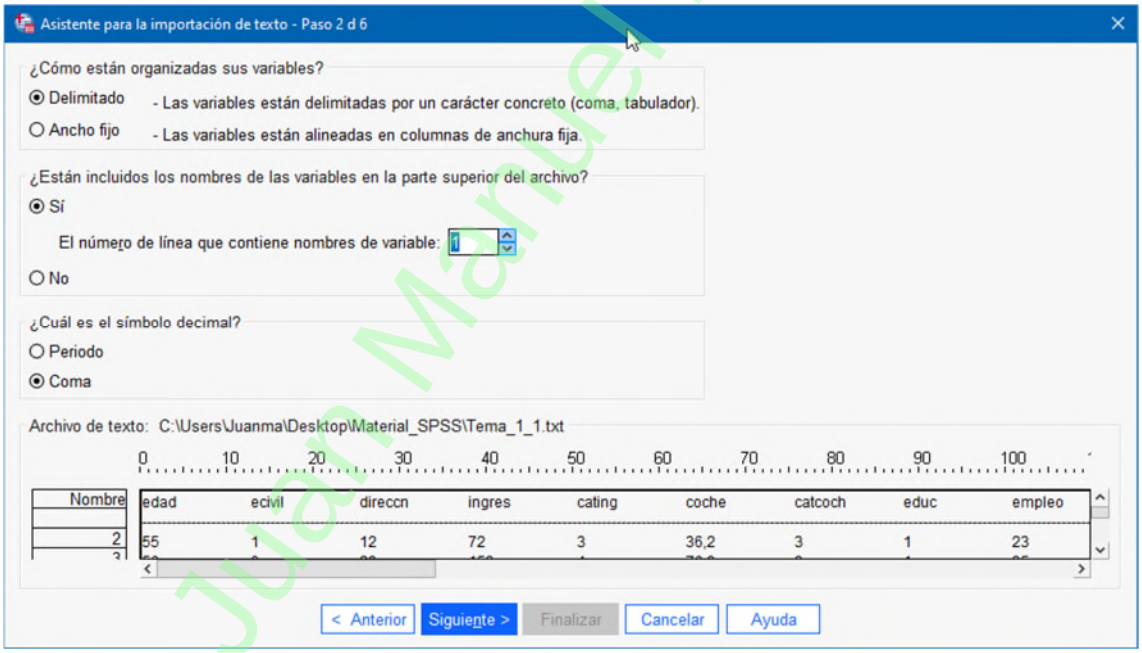


Elegir el tipo "Texto (*.txt, *.dat)" de manera que aparecerán los archivos con esa extensión (si los hay).

Siga los pasos indicados en el Asistente para la importación de texto para definir cómo desea leer el archivo de datos de texto:



El archivo de texto se mostrará en una ventana de vista previa. Puede aplicar un formato predefinido (guardado con anterioridad desde el Asistente para la importación de texto) o seguir los pasos del asistente para especificar cómo desea que se lean los datos.



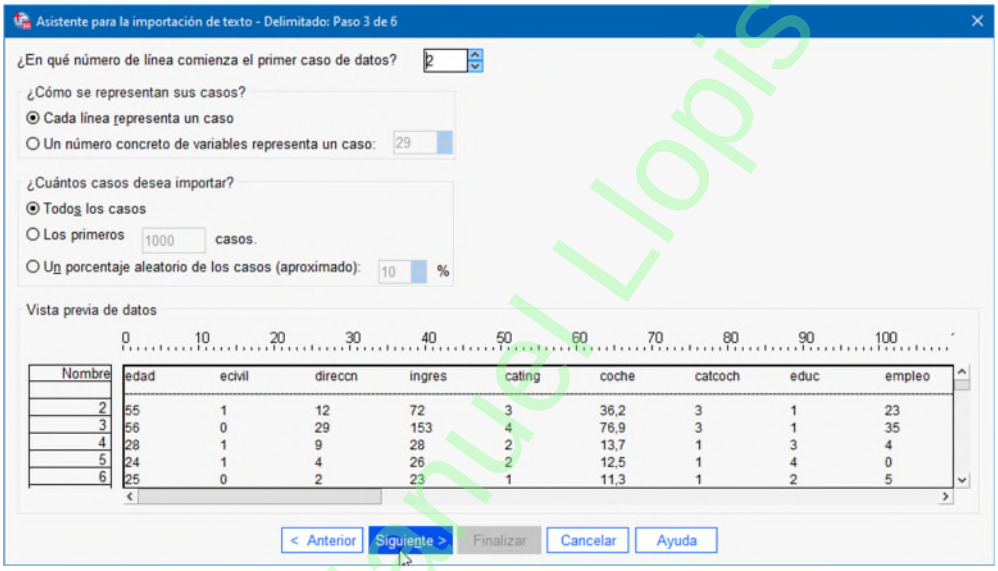
Este paso ofrece información sobre las variables. Por ejemplo, cada elemento de un cuestionario es una variable.

¿Cómo están organizadas sus variables? Para leer los datos adecuadamente, el Asistente para la importación de texto necesita saber cómo determinar el lugar en el que terminan los valores de datos de una variable y comienzan los valores de datos de la variable siguiente. La organización de las variables define el método utilizado para diferenciar una variable de la siguiente.

Delimitado. Se utilizan espacios, comas, tabulaciones u otros caracteres para separar variables. Las variables quedan registradas en el mismo orden para cada caso, pero no necesariamente conservando la misma ubicación para las columnas.

Ancho fijo. Cada variable se registra en la misma posición de columna en el mismo registro (línea) para cada caso del archivo de datos. No se requiere delimitador entre variables De hecho, en muchos archivos de datos de texto generados por programas de ordenador, podría parecer que los valores de los datos se suceden, sin espacios que los separen. La ubicación de la columna determina qué variable se está leyendo.

¿Están incluidos los nombres de las variables en la parte superior del archivo? Si la primera fila del archivo de datos contiene etiquetas descriptivas para cada variable, podrá utilizar dichas etiquetas como nombres de las variables. Los valores que no cumplan las normas de denominación de variables se convertirán en nombres de variables válidos.



Este paso ofrece información sobre los casos. Por ejemplo, cada persona que responde a un cuestionario es un caso.

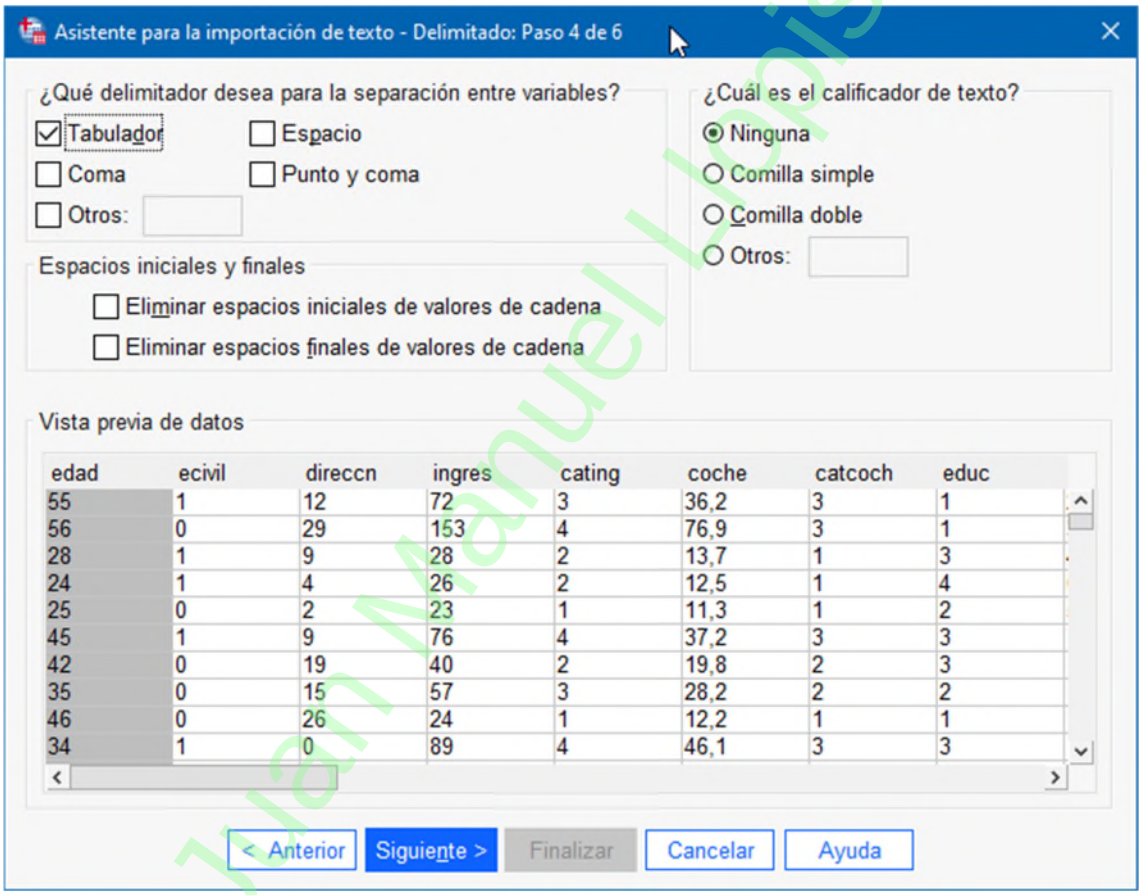
¿En qué número de línea comienza el primer caso de los datos? Indica la primera línea del archivo de datos que contiene valores de datos. Si la línea o líneas superiores del archivo de datos contienen etiquetas descriptivas o cualquier otro texto que no represente valores de datos, dicha línea o líneas no serán la línea 1.

¿Cómo se encuentran representados sus casos? Controla la manera en que el Asistente para la importación de texto determina dónde finaliza cada caso y comienza el siguiente:

Cada línea representa un caso. Cada línea contiene un sólo caso. Es bastante común que cada línea (fila) contenga un sólo caso, aunque dicha línea puede ser muy larga para un archivo de datos con un gran número de variables. Si no todas las líneas contienen el mismo número de valores de datos, el número de variables para cada caso quedará determinado por la línea que tenga el mayor número de valores de datos. A los casos con menos valores de datos se les asignarán valores perdidos para las variables adicionales.

Un número concreto de variables representa un caso. El número de variables especificado para cada caso informa al Asistente para la importación de texto de dónde detener la lectura de un caso y comenzar la del siguiente. Una misma línea puede contener varios casos y los casos pueden empezar en medio de una línea y continuar en la línea siguiente. El Asistente para la importación de texto determina el final de cada caso basándose en el número de valores leídos, independientemente del número de líneas. Cada caso debe contener valores de datos (o valores perdidos indicados por delimitadores) para todas las variables; de otra forma, el archivo de datos no se leerá correctamente.

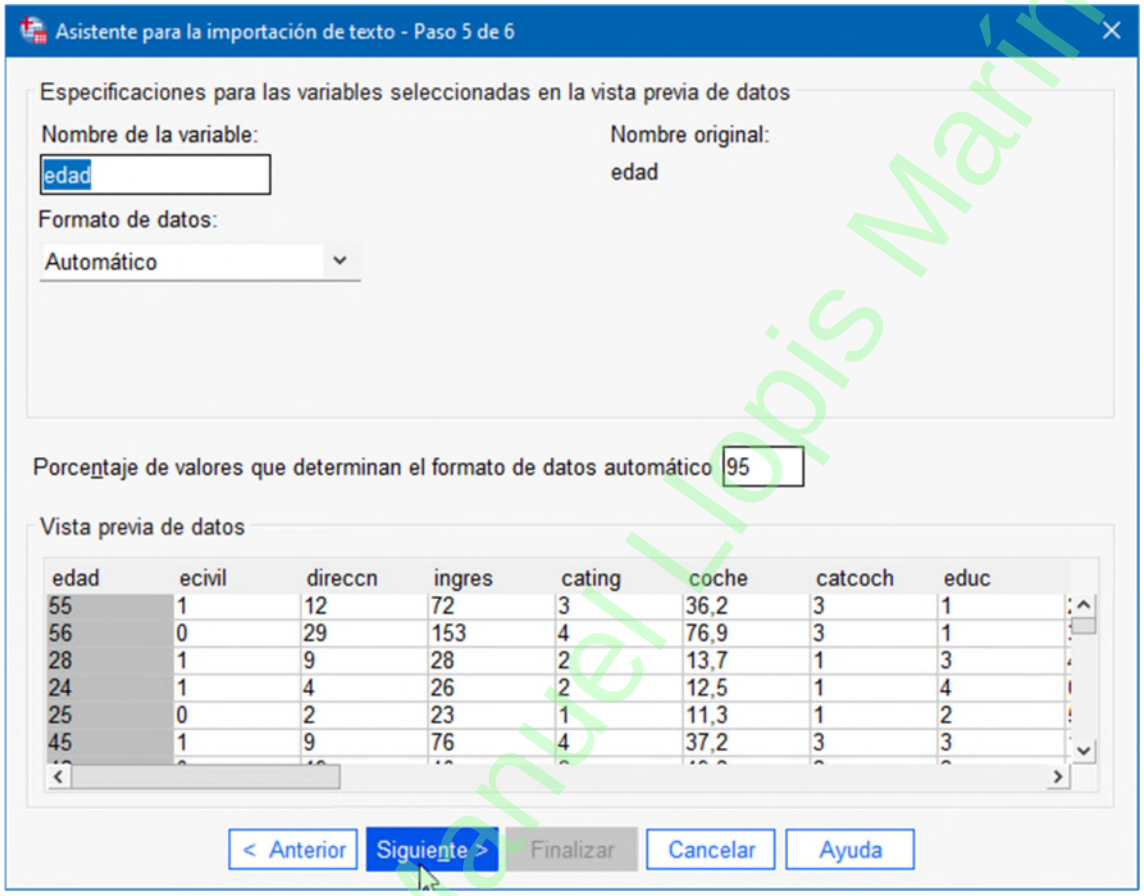
¿Cuántos casos desea importar? Puede importar todos los casos del archivo de datos, los primeros "n" casos (siendo n un número especificado por el usuario) o una muestra aleatoria a partir de un porcentaje especificado. Dado que esta rutina de muestreo aleatorio toma una decisión pseudoaleatoria para cada caso, el porcentaje de casos seleccionados sólo se puede aproximar al porcentaje especificado. Cuantos más casos contenga el archivo de datos, más se acercará el porcentaje de casos seleccionados al porcentaje especificado.



Este paso muestra la mejor opción, según el Asistente para la importación de texto, para leer el archivo de datos y le permite modificar la manera en que el asistente leerá las variables del archivo de datos.

¿Qué delimitador se encuentra entre las variables? Indica los caracteres o símbolos que separan los valores de datos. Puede seleccionar cualquier combinación de espacios, comas, signos de punto y coma, tabulaciones o cualquier otro carácter. En caso de existir varios delimitadores consecutivos sin valores de datos, dichos delimitadores serán considerados valores perdidos.

¿Cuál es el calificador de texto? Caracteres utilizados para encerrar valores que contienen caracteres delimitadores. Por ejemplo, si una coma es el delimitador, los valores que contengan comas se leerán incorrectamente a menos que estos valores se encierren en un calificador de texto, impidiendo que las comas del valor se interpreten como delimitadores entre los valores. Los archivos de datos con formato CSV de Excel utilizan las comillas dobles (") como calificador de texto. El calificador de texto aparece tanto al comienzo como al final del valor, encerrándolo completamente.



Este paso controla el nombre de la variable y el formato de datos que el Asistente para la importación de texto utilizará para leer cada variable, así como las que se incluirán en el archivo de datos definitivo.

Nombre de variable. Puede sobrescribir los nombres de variable predeterminados y sustituirlos por otros diferentes. Si lee nombres de variable desde el archivo de datos, el Asistente para la importación de texto modificará de manera automática los nombres de variable que no cumplan las normas de denominación de variables.

Seleccione una variable en la ventana de vista previa e introduzca un nombre de variable.

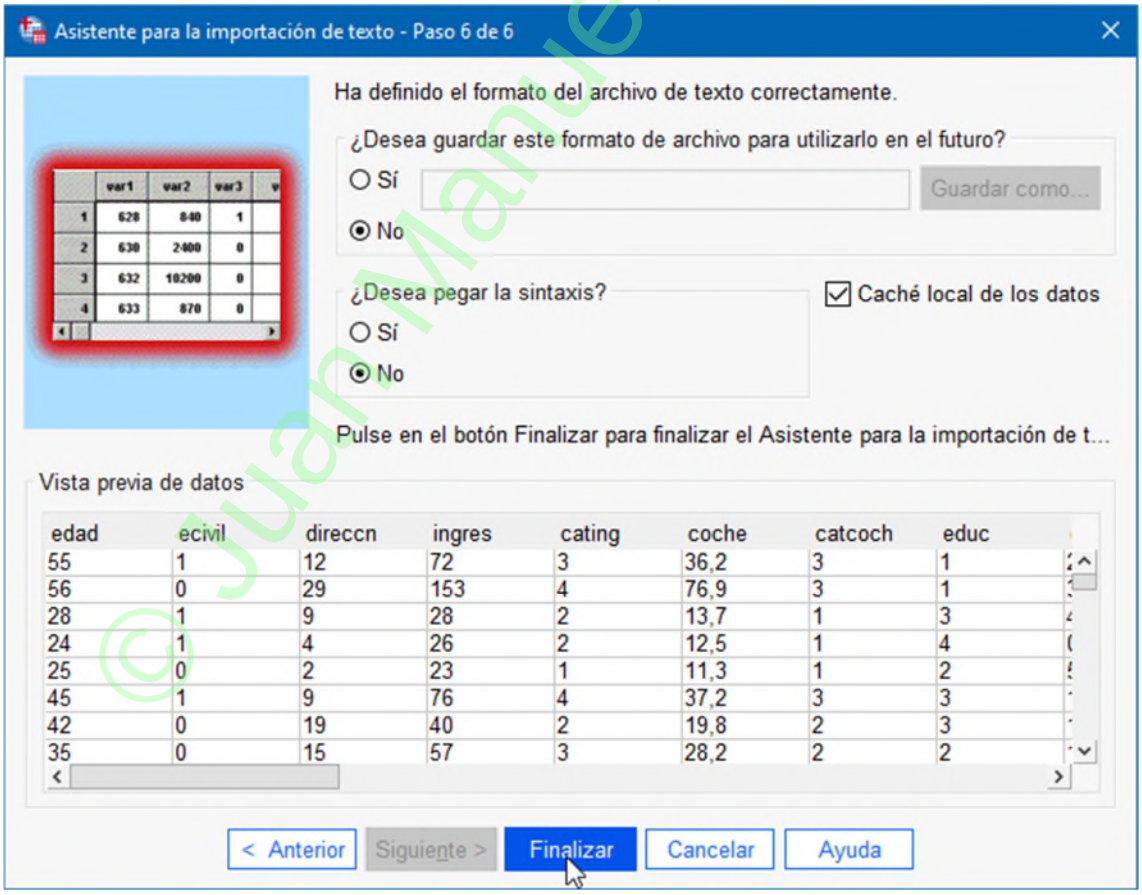
Formato de datos. Seleccione una variable en la ventana de vista previa y, a continuación, seleccione un formato de la lista desplegable.

Opciones para el formato de datos

Entre las opciones de formato para la lectura de variables con el Asistente para la importación de texto se encuentran:

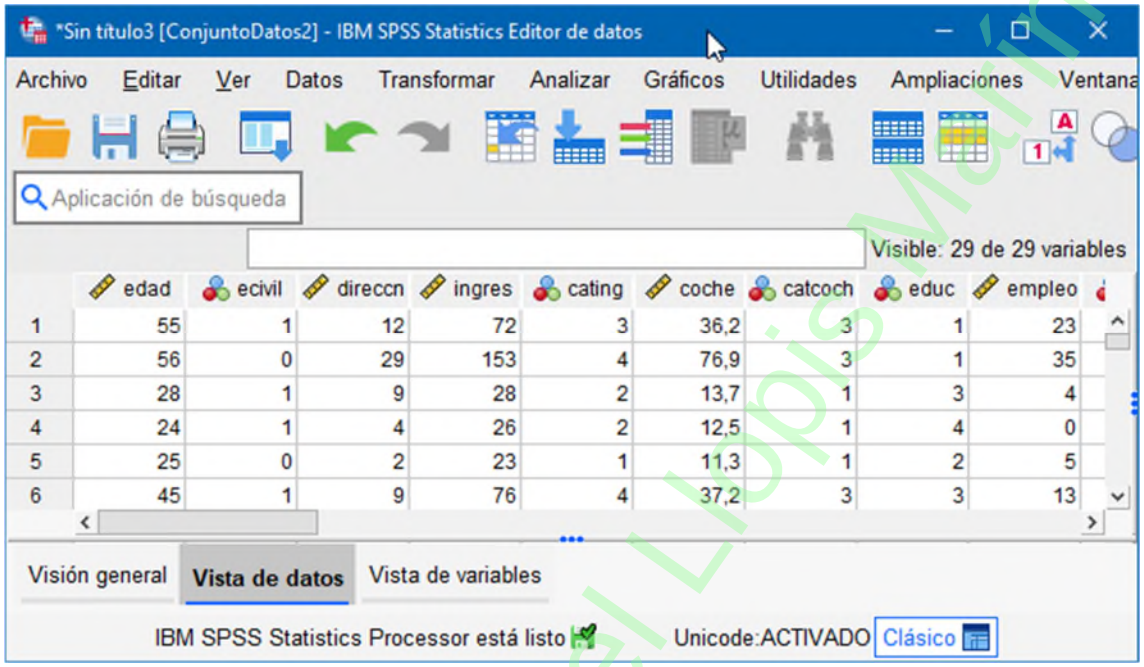
No importar. Omite la variable o variables seleccionadas del archivo de datos importado.
Numérico. Los valores válidos incluyen números, los signos más y menos iniciales y un indicador decimal.
Cadena. Son valores válidos prácticamente todos los caracteres del teclado y los espacios en blanco incrustados. En los archivos delimitados, puede especificar hasta un máximo de 255 de caracteres para el valor. El Asistente para la importación de texto fija como valor predeterminado para el número de caracteres el valor de cadena más largo que se haya encontrado para la variable o variables seleccionadas. Para los archivos de ancho fijo, el número de caracteres en los valores de cadena queda definido por la ubicación de las líneas de ruptura de variable en el paso 4. Defina el número de caracteres en el cuadro que aparece a la derecha.
Fecha/hora. Entre los valores válidos se encuentran las fechas con formato general: *dd-mm-aaaa*, *mm/dd/aaaa*, *dd.mm.aaaa*, *aaaa/mm/dd*, *hh:mm:ss*, así como una amplia variedad de formatos de hora y fecha. Los meses se pueden representar con dígitos, números romanos, abreviaturas de tres letras o con el nombre completo. Seleccione un formato de fecha de la lista que aparece a la derecha.
Dólar. Los valores válidos son números con un signo dólar inicial optativo y puntos separadores de millares también optativos.
Coma. Entre los valores válidos se encuentran los números que utilizan un punto para separar los decimales y una coma para separar los millares.
Puntos. Entre los valores válidos se encuentran los números que utilizan una coma para separar los decimales y un punto para separar los millares.

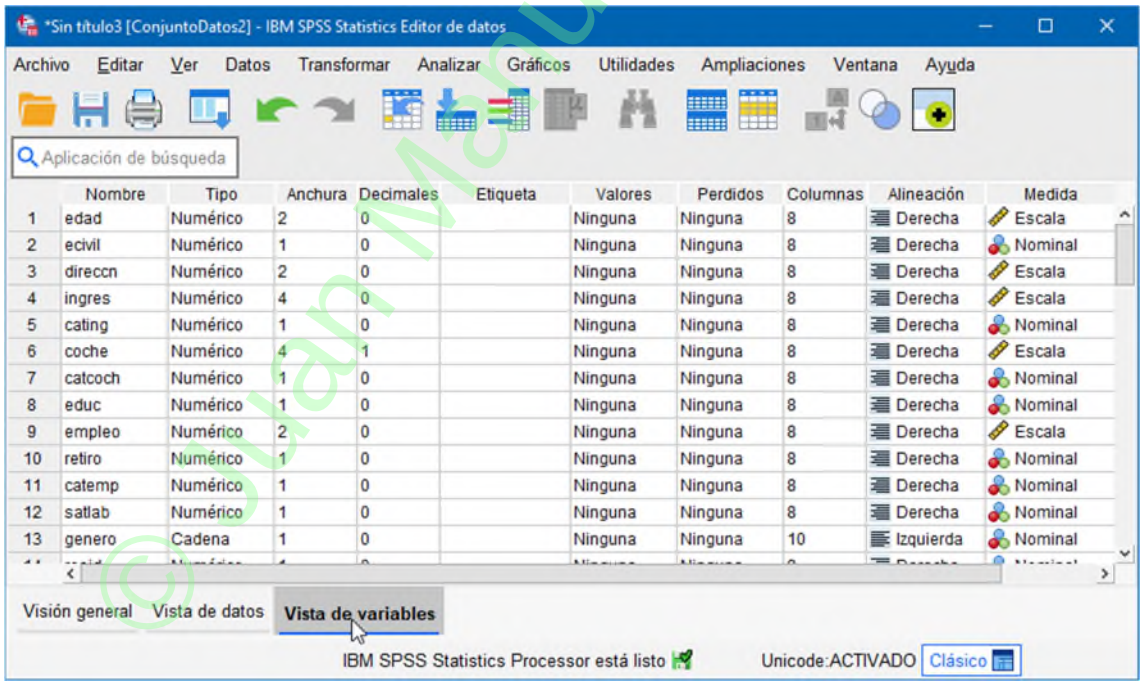
Nota: Los valores que contengan caracteres no válidos para el formato seleccionado serán considerados valores perdidos.



Este es el paso final del Asistente para la importación de texto. Puede guardar sus propias especificaciones en un archivo para hacer uso de ellas cuando importe archivos de datos de texto similares. También puede pegar la sintaxis generada por el Asistente para la importación de texto en una ventana de sintaxis. Así podrá personalizar y/o guardar dicha sintaxis para utilizarla en futuras sesiones o en trabajos de producción.

Al pulsar el botón "Finalizar", los datos pasan a la ventana de introducción de datos:

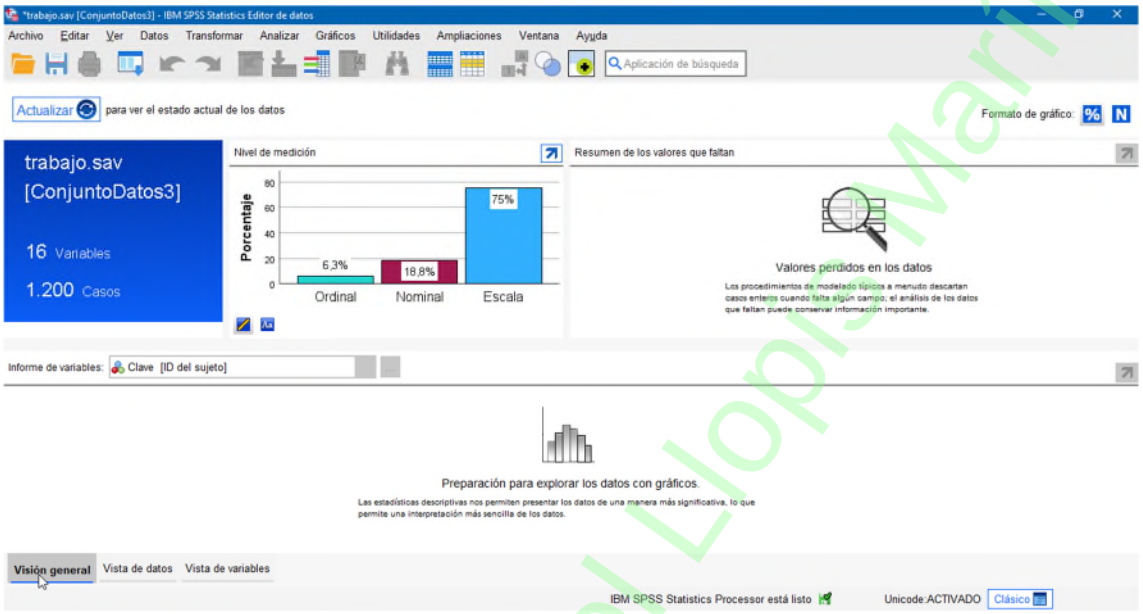




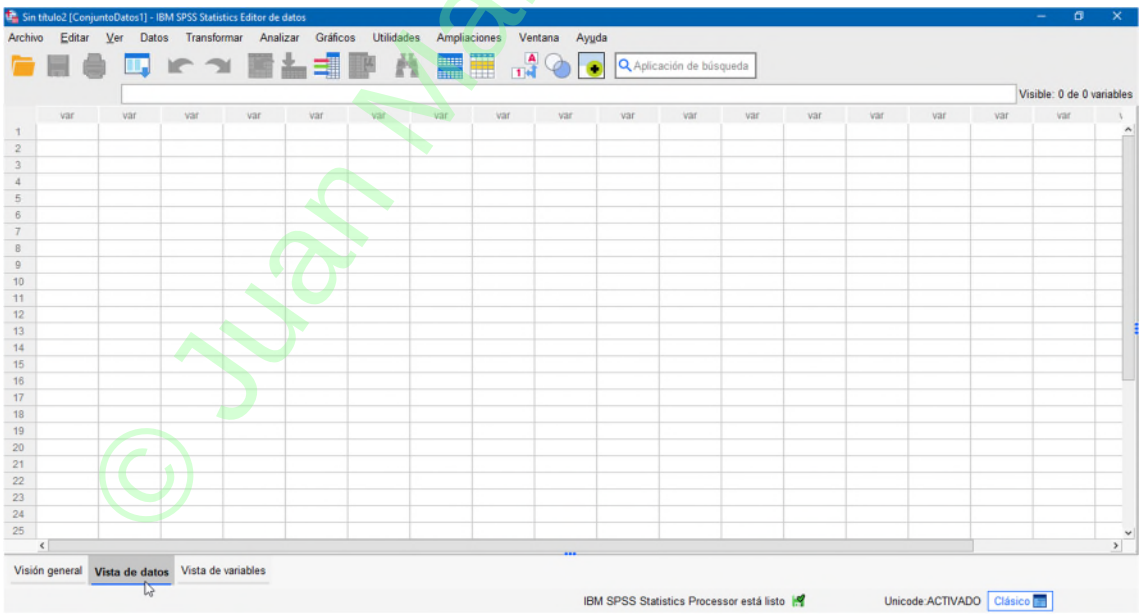
2.- Definición de variables

En la figura siguiente podemos ver el aspecto de la ventana de introducción de datos. En la parte inferior de la misma aparecen tres pestañas:

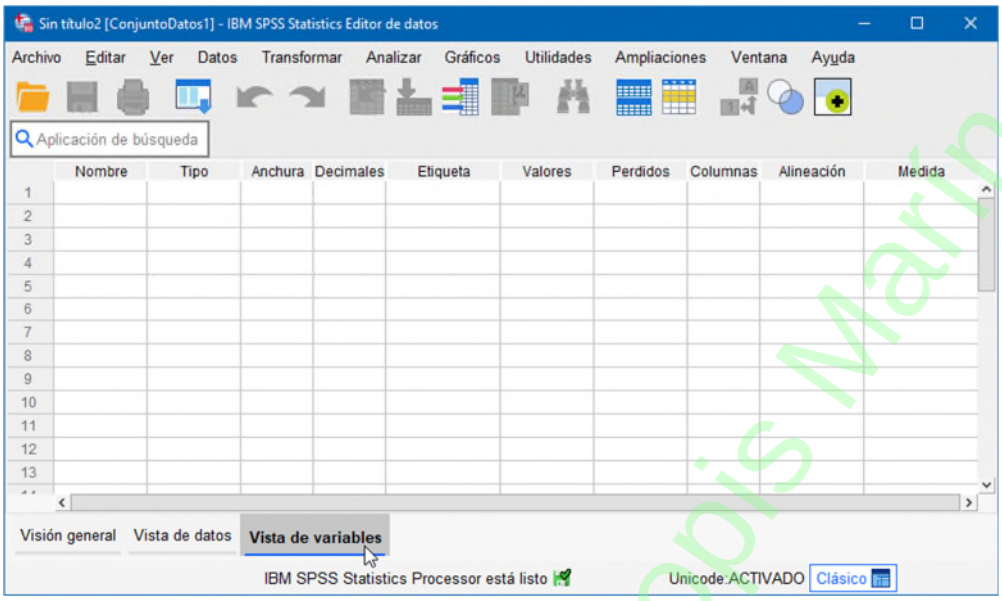
- La primera (**visión general**) presenta dos secciones: La sección superior proporciona información sobre un conjunto de datos o archivo completo. La sección inferior muestra información básica sobre una variable seleccionada. Sólo podrá seleccionarse cuando tengamos datos introducidos en la pestaña "vista de datos".



- La segunda **(vista de datos)** pertenece a la ventana de datos propiamente dicha, en donde los datos se introducen y manipulan de la misma manera y con las mismas opciones de copiado, etc. que en cualquier otro programa del sistema operativo Windows.

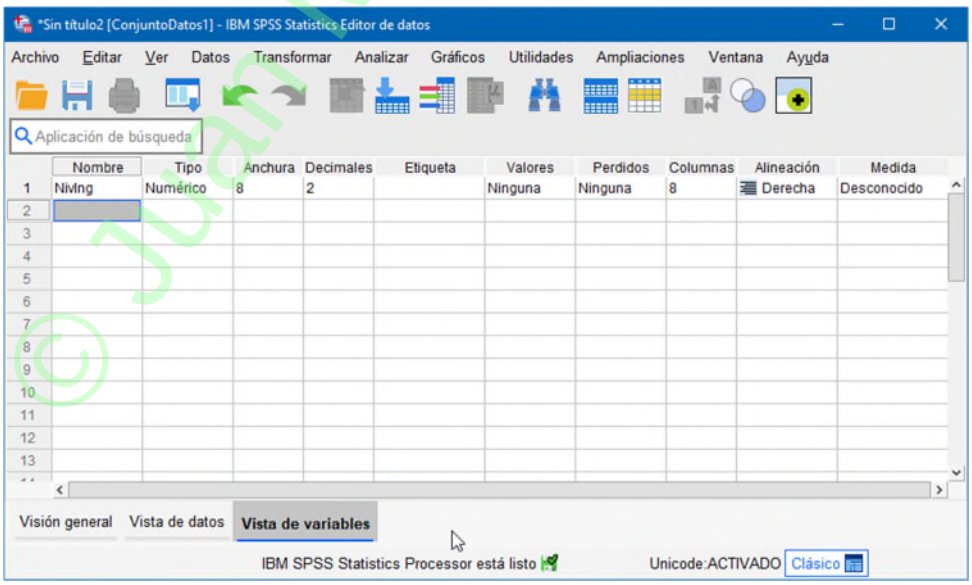


- La tercera pestaña **(vista de variables)** es más interesante desde el punto de vista del SPSS ya que nos da acceso a una hoja de captura de datos en la cual definiremos todas las variables que vamos a utilizar, así como sus características principales: métrica, tipo, número de decimales, etc.

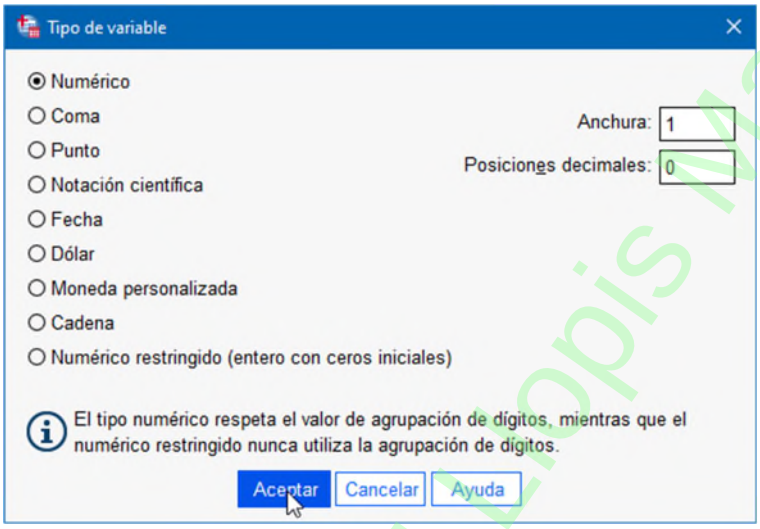


Dentro ya de la vista de variables nos encontramos con la posibilidad de definir las distintas características de nuestras variables. Así por ejemplo podemos introducir el nombre de la primera variable "NivIng" que hace referencia al nivel de ingresos de un grupo de sujetos. Debemos tener en cuenta que la extensión del nombre de la variable, a partir de la versión 12 de SPSS, ha aumentado de 8 caracteres a 64, pero sin espacios en blanco ni caracteres especiales (por ejemplo, !, ?, ' y *). (Podríamos haber introducido como nombre de variable: "Nivel_de_Ingresos").

SPSS por defecto nos definirá dicha variable tal y como lo vemos en la siguiente figura, es decir, numérica, con 8 dígitos y dos decimales, sin etiquetas ni de variable ni de valor, sin definición de datos perdidos, visualizando ocho dígitos, alineando los datos a la derecha y sin definir el nivel de medida de la variable.



Veamos en que forma definiríamos dichos valores para nuestro caso concreto y las opciones que nos da el programa. En primer lugar, pulsamos con el ratón sobre la palabra numérico lo cual provoca que se abra una ventana de captura de datos en la que definimos tanto el tipo de variable como el número máximo de dígitos. Cabe señalar que entre todas las opciones que nos presenta las más interesantes son "Numérica" y "Cadena". La primera como es obvio hace referencia a aquellas variables que son cifras y que no precisan de ningún tipo de presentación especial (delimitadores, signos monetarios, etc.) mientras que cadena hace referencia a variables que no son tratadas de forma numérica, como puede ser el nombre del sujeto, etc.



Puede observarse que en esta ventana hemos definido también una anchura de un dígito y ningún decimal, dado que la variable NivIng, solo presentará un dígito y ningún decimal.

En la columna **etiqueta** podemos introducir una definición más amplia de la variable de tal modo que en próximas ocasiones podamos saber a qué refiere, en nuestro caso hemos introducido "Nivel de Ingresos", esta etiqueta será también visible en las ventanas de captura de datos de los análisis que realicemos con posterioridad. Es preciso señalar la importancia de definir este tipo de etiquetas en las versiones de SPSS anteriores a la versión 12, dado que la limitación a ocho caracteres de las variables hacía que, en muchos casos, al cabo de un tiempo no supiéramos a qué hacían referencia. En la versión 12, solamente tiene sentido si no queremos "sobrecargar" la ventana de vista de datos.

La siguiente columna; **valores** (**son las etiquetas de valor)**, nos permite definir etiquetas particulares para cada valor de la variable. La utilidad de esta opción aparece cuando trabajamos con variables categóricas como la del ejemplo. Así la variable "Nivel de Ingresos" la vamos a codificar de la siguiente manera:

1 – Muy Bajo.

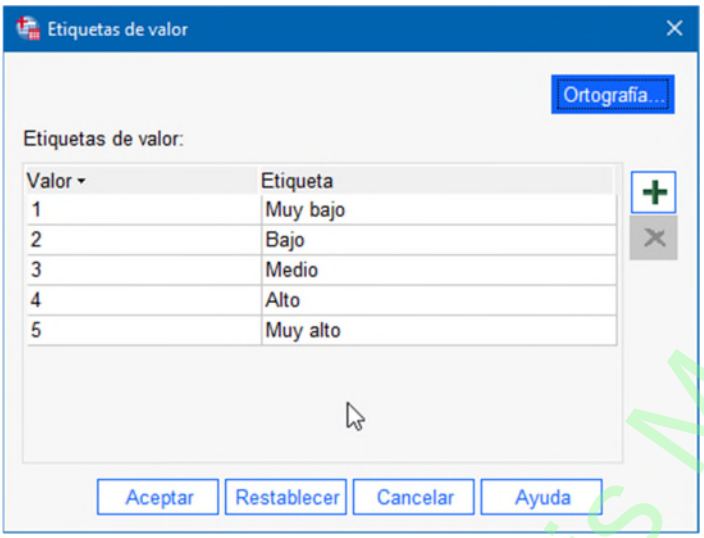
2 – Bajo

3 – Medio

4 – Alto

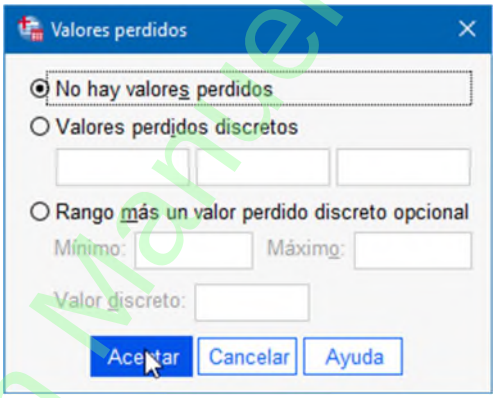
5 – Muy Alto

Para ello pulsaremos sobre la casilla correspondiente y nos aparecerá una ventana de captura de datos como la siguiente:



En esta ventana introducimos el valor 1, la etiqueta Muy Bajo y a continuación **pulsamos el signo +**, posteriormente hacemos lo mismo para los siguientes valores y finalizamos pulsando *Aceptar.*

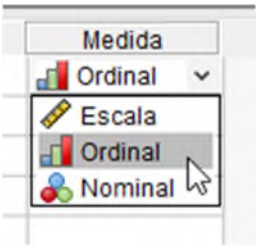
A continuación, definimos los **valores perdidos**, es decir, aquellos valores de nuestra variable que representan los casos que, o bien no hemos obtenido, o bien no son válidos. La ventana correspondiente puede observarse en la imagen inferior.



Esta opción tan solo será necesaria en aquellos casos en que los datos de que disponemos hayan sido introducidos con un valor perdido concreto, dado que, si introducimos los datos nosotros siempre tenemos la posibilidad de no introducir ningún valor en aquella casilla que contenga un dato perdido, con lo que SPSS lo reconocerá así, aunque no efectuemos ninguna definición de estos. Podemos comprobar que además de introducir valores concretos, SPSS nos da la posibilidad de definir un rango de valores como valores perdidos.

Las dos opciones siguientes; **columnas** y **alineación**, se refieren a la visualización de las variables en la ventana "vista de datos" y modifican el número de columnas que se visualizan, así como la alineación (derecha, izquierda o centro) de los datos.

Finalmente podemos introducir la **métrica** en que está medida la variable, en nuestro caso al ser categórica, pero ordenable, quedaría comprendida en la categoría "ordinal". Para ello utilizamos la ventana de captura de datos que vemos a continuación.



Una advertencia muy importante es que, aunque nosotros definamos la métrica de la variable, SPSS no nos va a advertir, en un análisis concreto, de la inadecuación de este en función de la métrica de las variables. De este modo podemos ver cómo es posible solicitar una regresión lineal con dos variables nominales y SPSS ejecutará la misma sin generar ningún error, lo cual suele llevar a interesantes trabajos en los que se demuestra la relación entre el color del pelo y el sexo de los ángeles. Hay que tener siempre en cuenta el hecho de que si el programa no nos da ningún error no implica que no estemos equivocándonos totalmente.

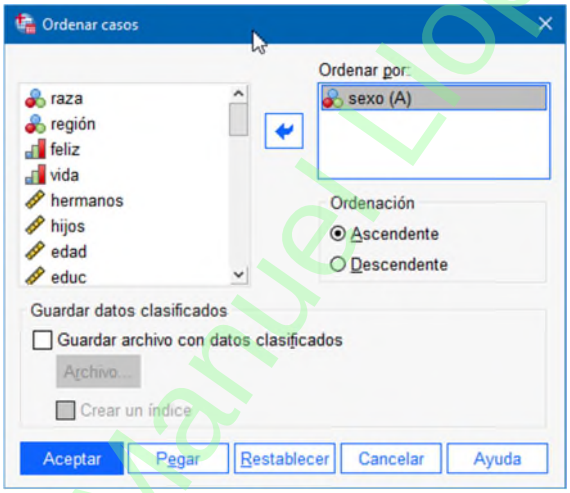
TEMA - 2

Ordenación, selección y ponderación de casos

1.- Ordenación de casos

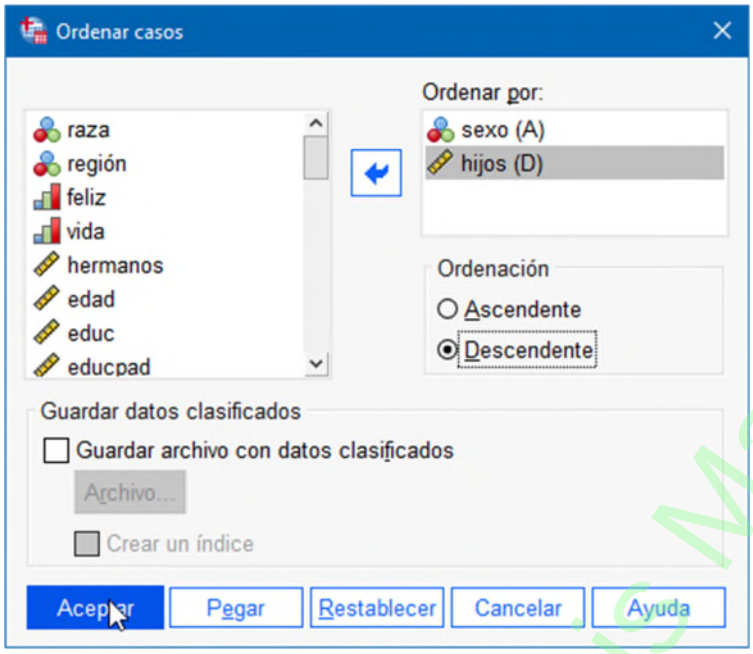
La finalidad de esta manipulación es la de ordenar de modo creciente o decreciente a los sujetos en función de las puntuaciones en alguna de las variables. Si, por ejemplo, detectamos que al calcular la Media Aritmética de una variable se obtiene un resultado absurdo (mucho mayor que el máximo de la escala, por ejemplo), es conveniente revisar los datos (ya que probablemente se ha cometido un error en la entrada de estos). Ordenando los sujetos de mayor a menor en función de la variable estudiada, el sujeto con la puntuación mayor en dicha variable quedará ordenado en primer lugar (pudiendo rectificar ahora su puntuación). Esta opción también es muy útil en el caso de que debamos presentar un listado de los casos con los que estamos trabajando ordenado en función de alguna o algunas de las variables incluidas en nuestros datos. **Fichero: encuesta.sav**

Seleccionando la opción *Ordenar casos…* en el menú *Datos*, se activa el siguiente submenú:



En este submenú se ha seleccionado la variable "sexo" para ser ordenada de modo *Ascendente* (de menor a mayor). Al pulsar el botón *Aceptar*, los casos se reordenarán en la ventana de datos en función de dicha variable.

Por otra parte, es posible ordenar los casos en función de dos o más criterios. De este modo los casos quedarán ordenados en función de la primera variable seleccionada, los casos con un mismo valor en dicha variable quedarán ordenados en función del segundo criterio y así sucesivamente. Así, por ejemplo, podríamos hacer un listado en el que aparecieran en primer lugar los hombres y después las mujeres, estando a su vez ordenados dentro de cada género en función de su número de hijos (variable "hijos") pero en esta variable, de mayor a menor, de forma descendente. La ordenación de cada variable, desde hace algunas versiones, es independiente de las otras.

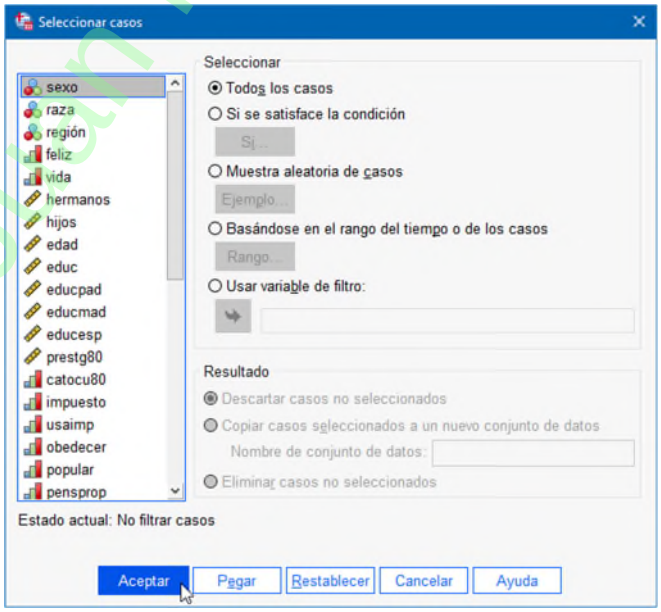


2.- Selección de casos

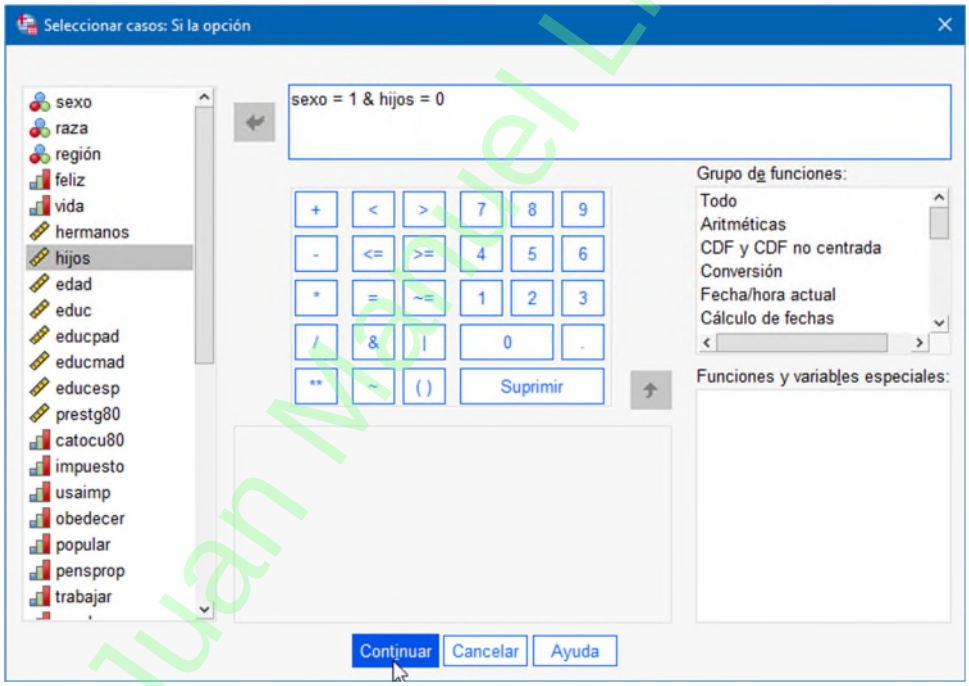
Seleccionar casos proporciona varios métodos para seleccionar un subgrupo de casos basándose en criterios que incluyen variables y expresiones complejas. También se puede seleccionar una muestra aleatoria de casos. Los criterios usados para definir un subgrupo pueden incluir:

- Valores y rangos de las variables
- Rangos de fechas y horas
- Números de caso (filas)
- Expresiones aritméticas
- Expresiones lógicas
- Funciones

Seleccionando la opción *Seleccionar casos…* en el menú *Datos*, se activa el siguiente submenú:

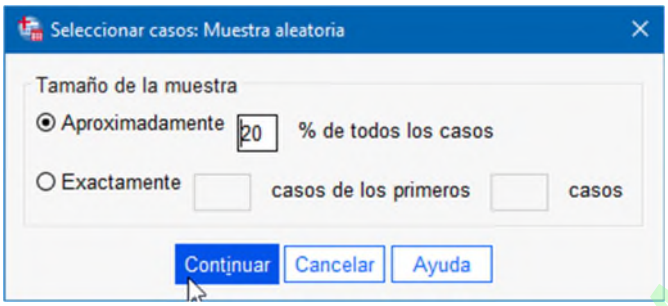


Todos los casos. Desactiva el filtrado de casos y utiliza todos los casos.
Si se satisface la condición. Utiliza una expresión condicional para seleccionar los casos. Si el resultado de la expresión condicional es verdadero, el caso se selecciona. Si el resultado es falso o perdido, entonces el caso no se selecciona.
Muestra aleatoria de casos. Selecciona una muestra aleatoria basándose en un porcentaje aproximado o en un número exacto de casos.
Basándose en el rango del tiempo o de los casos. Selecciona los casos basándose en un rango de los números de caso o en un rango de las fechas/horas.
Usar variable de filtro. Utiliza como variable para el filtrado la variable numérica seleccionada del archivo de datos. Se seleccionan los casos con cualquier valor distinto del 0 o del valor perdido para la variable seleccionada.
Casos no seleccionados. Puede filtrar o eliminar los casos que no reúnan los criterios de selección. Los casos filtrados permanecen en el archivo de datos, pero se excluyen del análisis. Seleccionar casos crea una variable de filtro, "filter_$", para indicar el estado del filtro. Los casos seleccionados tienen un valor de 1; los casos filtrados tienen un valor de 0. Estos últimos también se indican con una barra transversal sobre el número de fila en el Editor de datos. Para desactivar el filtrado e incluir todos los casos en el análisis, seleccione Todos los casos. Si guarda el archivo de datos después de eliminar casos, no podrá recuperar los casos eliminados.
Seleccionar casos: Si
Este cuadro de diálogo permite seleccionar subconjuntos de casos utilizando expresiones condicionales. Las expresiones condicionales devuelven un valor verdadero, falso o perdido para cada caso.



- Si el resultado de una expresión condicional es *verdadero*, el caso se selecciona.
- Si el resultado de una expresión condicional es *falso* o *perdido*, no se selecciona el caso.
- La mayoría de las expresiones condicionales utilizan al menos uno de los seis operadores de relación (<, >, <=, >=, =, y ~=) en el teclado de calculadora.
- Las expresiones condicionales pueden incluir nombres de variable, constantes, operadores aritméticos, funciones numéricas y otras, variables lógicas y operadores relacionales.

Seleccionar casos: Muestra aleatoria
Este cuadro de diálogo permite seleccionar una muestra aleatoria basada en un porcentaje aproximado o en un número exacto de casos. El muestreo se realiza sin sustitución, de manera que el mismo caso no se puede seleccionar más de una vez.



Aproximadamente. Genera una muestra aleatoria con el porcentaje aproximado de casos indicado. Dado que esta rutina toma una decisión pseudoaleatoria para cada caso, el porcentaje de casos seleccionados sólo se puede aproximar al especificado. Cuantos más casos contenga el archivo de datos, más se acercará el porcentaje de casos seleccionados al porcentaje especificado.
Exactamente. Un número de casos especificado por el usuario. También se debe especificar el número de casos a partir de los cuales se generará la muestra. Este segundo número debe ser menor o igual que el número total de casos presentes en el archivo de datos. Si lo excede, la muestra contendrá un número menor de casos proporcional al número solicitado.

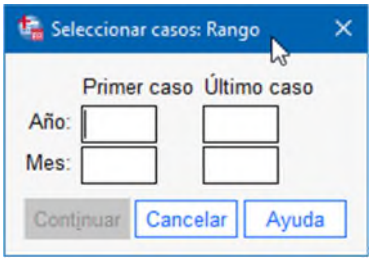
Seleccionar casos: Rango
Este cuadro de diálogo selecciona los casos basándose en un rango de números de caso o en un rango de fechas u horas.

- Los rangos de casos se basan en el número de fila que se muestra en el Editor de datos.

- Los rangos de fechas y horas sólo están disponibles para **datos de series temporales** con variables de fecha definidas (menú Datos, Definir fechas).

Submenú para una variable que no sea de fecha:

Submenú para datos de una serie temporal con variables de fecha definidas como años y meses (menú Datos --- Definir fechas):



3.- Ponderación de casos

Cuando cada registro de los que componen un fichero de trabajo representa más de un caso, SPSS permite especificar el coeficiente de ponderación correspondiente.

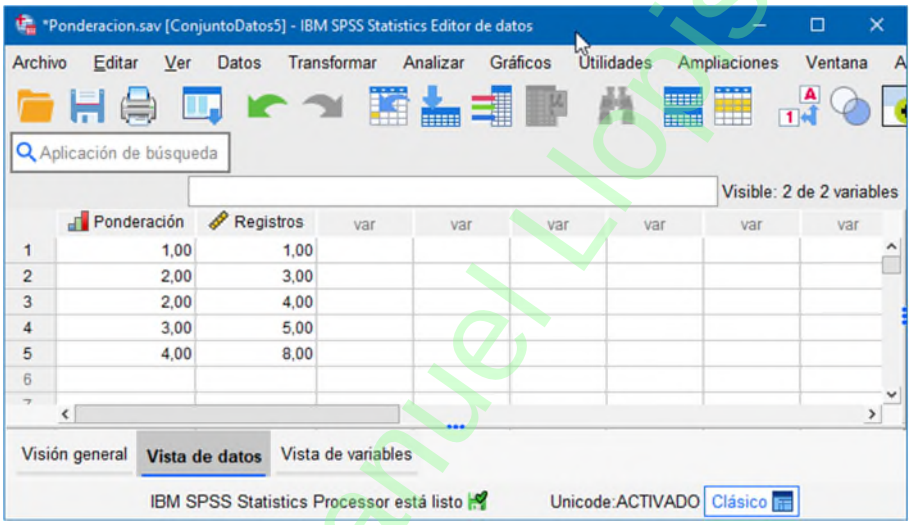
Es requisito que en el fichero exista una variable cuyos valores van a ser empleados como coeficientes de ponderación.

Ponderar casos proporciona a los casos diferentes pesos (mediante una réplica simulada) para el análisis estadístico.

Los valores de la variable de ponderación deben indicar el número de observaciones representado por los casos únicos del archivo de datos.

Los casos con valores perdidos, negativos o cero para la variable de ponderación se excluyen del análisis.

Ejemplo: Fichero: ponderación.sav

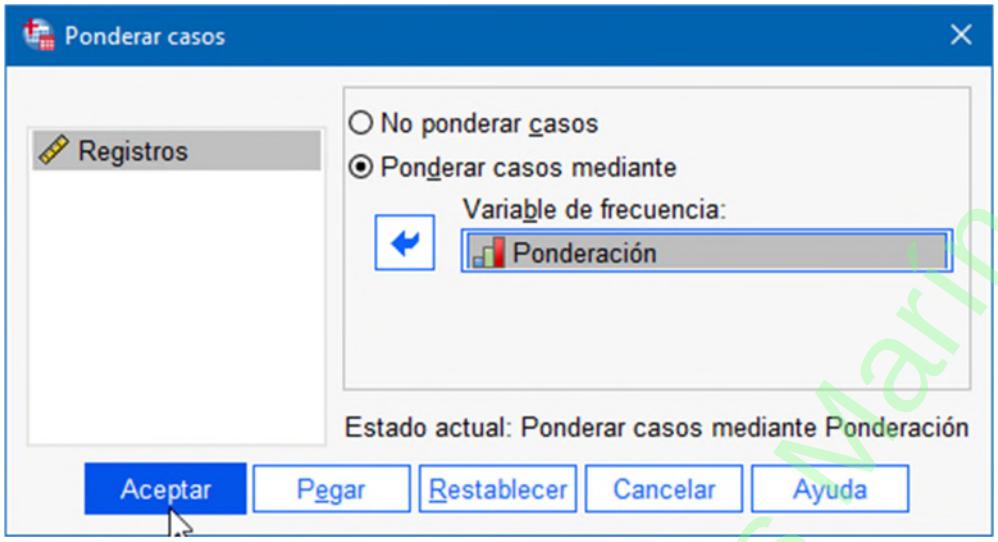


Frecuencias de la variable "Registros" sin ponderar:

Registros

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1,00	1	20,0	20,0	20,0
	3,00	1	20,0	20,0	40,0
	4,00	1	20,0	20,0	60,0
	5,00	1	20,0	20,0	80,0
	8,00	1	20,0	20,0	100,0
	Total	5	100,0	100,0	

Frecuencias de la variable “Registros” ponderada según la variable “Ponderación”:



Registros

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1,00	1	8,3	8,3	8,3
	3,00	2	16,7	16,7	25,0
	4,00	2	16,7	16,7	41,7
	5,00	3	25,0	25,0	66,7
	8,00	4	33,3	33,3	100,0
	Total	12	100,0	100,0	

TEMA - 3

Transformación de datos: Cálculo (Creación de nuevas variables) y recodificación

1.- Cálculo (Creación de nuevas variables)

La creación de variables implica el cálculo de nuevas variables en función de las variables ya existentes, o bien en función de criterios externos a las variables ya existentes. Un ejemplo donde se requiere la creación de nuevas variables podría ser el siguiente: se tienen dos puntuaciones de cada sujeto (cada una en una variable) y se pretende calcular la puntuación promedio para cada sujeto en las dos variables. Calcular manualmente el promedio sujeto por sujeto es una solución muy tediosa. En estos casos SPSS dispone de comandos que permiten generar nuevas variables a partir de transformaciones de las existentes. **Fichero: calculo.sav**

Seleccionando la opción *Calcular* del menú *Transformar* se activa el Cuadro de Diálogo *Calcular Variable* mostrado a continuación:

Los aspectos que se deben considerar son:

a.- Escriba el nombre de una sola variable de destino. Puede ser una variable existente o una nueva que se vaya a añadir al archivo de datos de trabajo. Este nombre se especifica en la ventana *Variable de destino*.

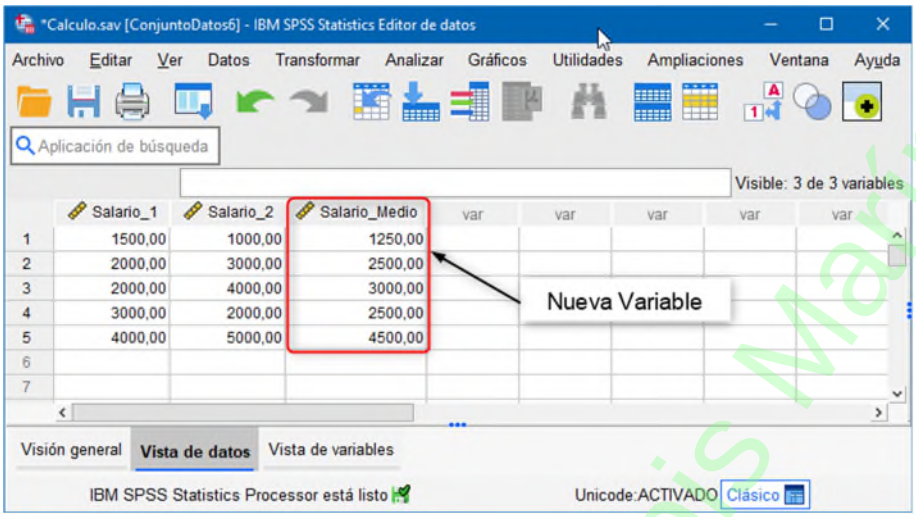
b.- El botón *Tipo y etiqueta* permite definir el tipo de la variable nueva, así como sus etiquetas. Para las nuevas variables de cadena, deberá seleccionar *Tipo y etiqueta* obligatoriamente (por defecto, SPSS considera las variables de tipo numérico.

c.- Las constantes de cadena deben ir entre comillas o apóstrofos.

d.- Las constantes numéricas deben escribirse en formato americano, con el punto (.) como separador decimal.

e.- En la ventana *Expresión numérica* se debe escribir la función que define la nueva variable. Aquí se pueden utilizar otras variables, valores constantes y las funciones que se muestran en la ventana *Funciones*. Pegue las funciones de la lista de funciones y rellene los parámetros indicados por signos de interrogación.

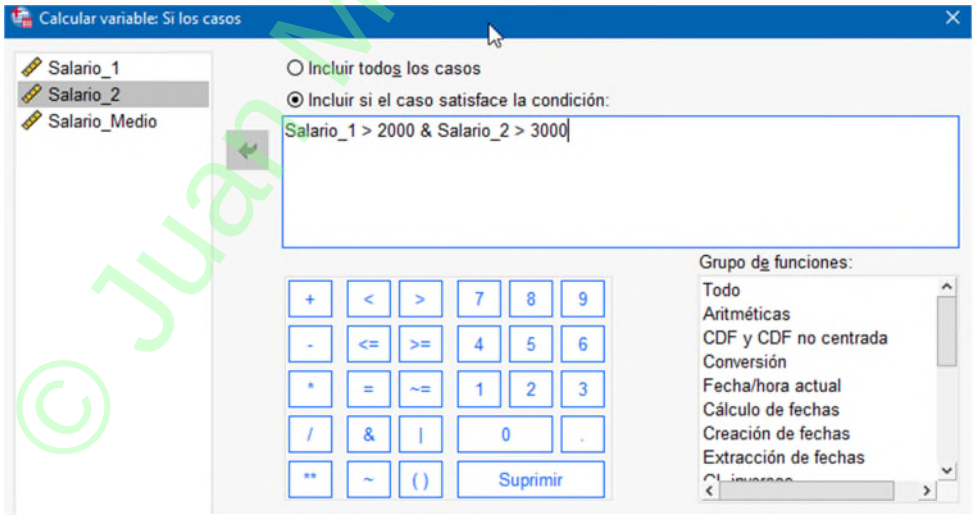
En la figura de la página anterior se está creando una nueva variable *Salario_Medio*, que será la media de las variables *Salario_1* y *Salario_2* de cada sujeto. Al pulsar el botón *Aceptar*, SPSS creará la nueva variable y se añadirá en la ventana de datos de SPSS, como se muestra a continuación.

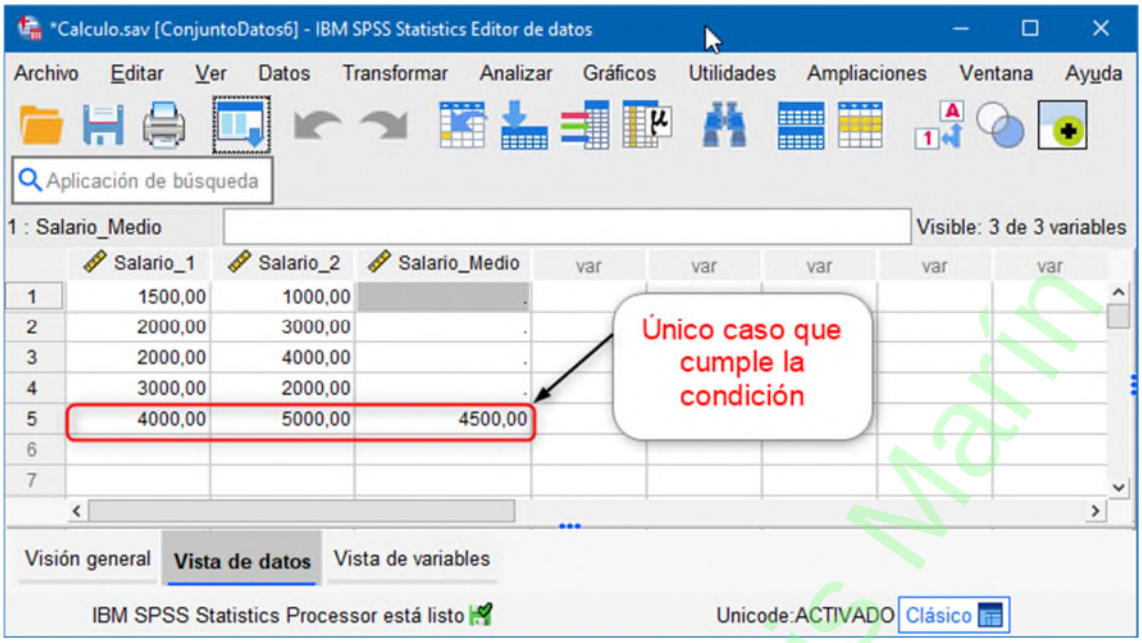


Calcular variable: Si los casos

El cuadro de diálogo "Si los casos" permite aplicar transformaciones de los datos para subconjuntos de casos seleccionados utilizando expresiones condicionales. Una expresión condicional devuelve un valor *verdadero*, *falso* o *perdido* para cada caso.

- Si el resultado de una expresión condicional es *verdadero*, la transformación se aplicará al caso.
- Si el resultado de una expresión condicional es *falso* o *perdido*, no se aplicará la transformación al caso.
- La mayoría de las expresiones condicionales utiliza al menos uno de los seis operadores de relación (<, >, <=, >=, =, y ~=) del teclado de calculadora.
- Las expresiones condicionales pueden incluir nombres de variable, constantes, operadores aritméticos, funciones numéricas y otras, variables lógicas y operadores relacionales.





Funciones

Se dispone de muchos tipos de funciones, entre ellos:

Funciones aritméticas.
Funciones estadísticas.
Funciones de cadena.
Funciones de fecha y hora.
Funciones de distribución.
Funciones de variables aleatorias.
Funciones de valores perdidos.

Para obtener una lista completa de las funciones, busque funciones en el índice del sistema de Ayuda en pantalla de SPSS. También puede pulsar con el botón derecho del ratón sobre una función seleccionada en la lista del cuadro de diálogo para obtener una descripción de dicha función.

Valores perdidos en funciones
Las funciones y las expresiones aritméticas sencillas tratan los valores perdidos de diferentes formas. En la expresión: (var1+var2+var3)/3, el resultado es el valor perdido si un caso tiene un valor perdido para cualquiera de las tres variables.

En la expresión: MEAN(var1, var2, var3) el resultado es el valor perdido sólo si el caso tiene valores perdidos para las tres variables.

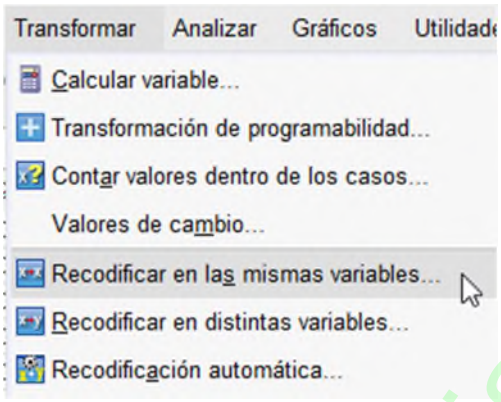
En las funciones estadísticas se puede especificar el número mínimo de argumentos que deben tener valores no perdidos. Para ello, escriba un punto y el número mínimo de argumentos después del nombre de la función, como en:

MEAN.2(var1, var2, var3)

2.- Recodificación de valores

La recodificación de datos implica el cambio de escala de alguna de las variables. Se pueden recodificar las variables numéricas y de cadena. Si selecciona múltiples variables, todas deben ser del mismo tipo. No se pueden recodificar juntas las variables numéricas y de cadena.

SPSS permite **recodificar** los valores **en la misma variable** que se recodifica o **en una nueva variable**. Es recomendable recodificar siempre en una nueva variable ya que, de este modo, la recodificación no implica la pérdida de los valores en la variable original. En el menú *Transformar* se encuentran las opciones posibles de recodificación.

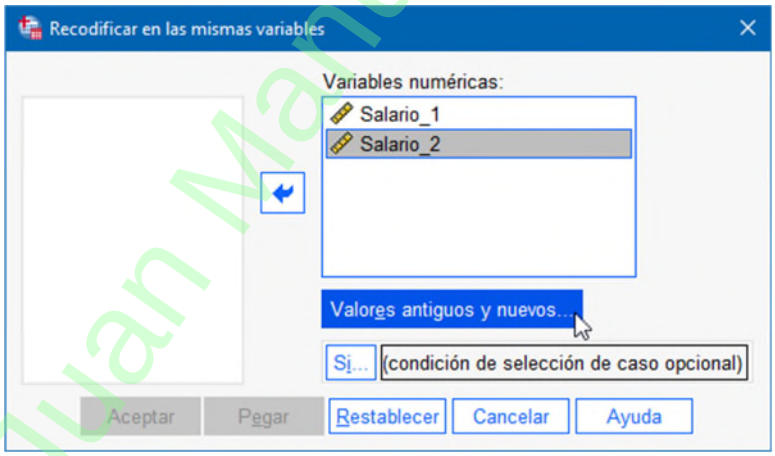


Recodificar en las mismas variables

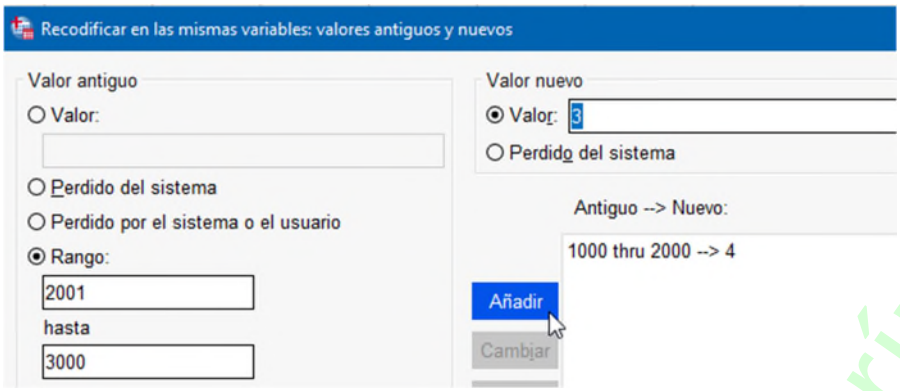
Recodificar en las mismas variables reasigna los valores de las variables existentes o agrupa rangos de valores existentes en nuevos valores en las mismas variables, sustituyendo los valores antiguos por los nuevos.

Para recodificar los valores de una variable, elija en los menús: Transformar --- Recodificar --- En las mismas variables.

Seleccione las variables que desee recodificar. Si selecciona múltiples variables, todas deberán ser del mismo tipo (numéricas o de cadena).



Pulse en Valores antiguos y nuevos y especifique cómo deben recodificarse los valores. El cuadro de diálogo que aparece permite definir los valores que se van a recodificar. Todas las especificaciones de valores deben pertenecer al mismo tipo de datos (numéricos o de cadena) que las variables seleccionadas en el cuadro de diálogo principal.



Valor antiguo. Determina el valor o los valores que se van a recodificar. Puede recodificar valores individuales, rangos de valores y valores perdidos. Los rangos y los valores perdidos del sistema no se pueden seleccionar para las variables de cadena, ya que ninguno de los conceptos es aplicable a estas variables. Los rangos incluyen sus puntos finales y los valores definidos como perdidos por el usuario que estén dentro del rango. **Elementos**:

Valor. Valor antiguo individual que se va a recodificar en un valor nuevo. El tipo de datos (numérico o de cadena) del valor introducido debe coincidir con el tipo de datos de las variables que se desea recodificar.

Perdido por el sistema. Valores asignados por SPSS cuando los valores de los datos no están definidos de acuerdo con el tipo de formato que se ha especificado, cuando un campo numérico está vacío o cuando no se ha definido un valor generado por un comando de transformación. Los valores numéricos perdidos del sistema se muestran como puntos. Las variables de cadena no pueden tener valores perdidos del sistema, ya que es lícito cualquier carácter en las variables de cadena.
Perdido por el sistema o usuario. Observaciones que tienen valores que, o bien han sido declarados como perdidos por el usuario o bien son desconocidos y han sido asignados el valor perdido del sistema, lo cual se indica mediante un punto (.).

Rango. Un rango inclusivo de valores. No está disponible para variables de cadena. Se incluirán cualesquiera valores perdidos definidos por el usuario que se encuentren dentro del rango.

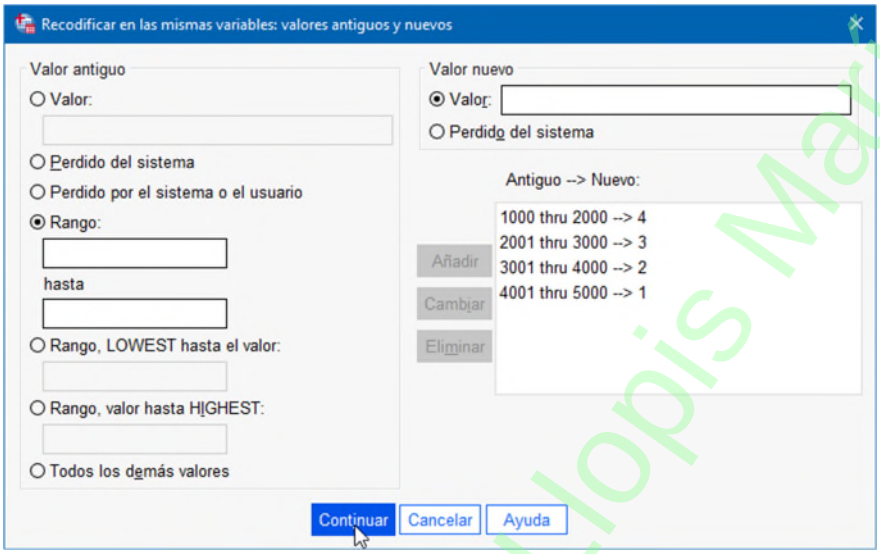
Todos los demás valores. Cualquier valor no incluido en una de las especificaciones de la lista Antiguo-Nuevo. Aparece en la lista Antiguo-Nuevo como ELSE.

Valor nuevo. Es el valor individual en el que se recodifica cada valor o rango de valores antiguos. Puede introducir un valor o asignar el valor perdido del sistema. **Elementos**:

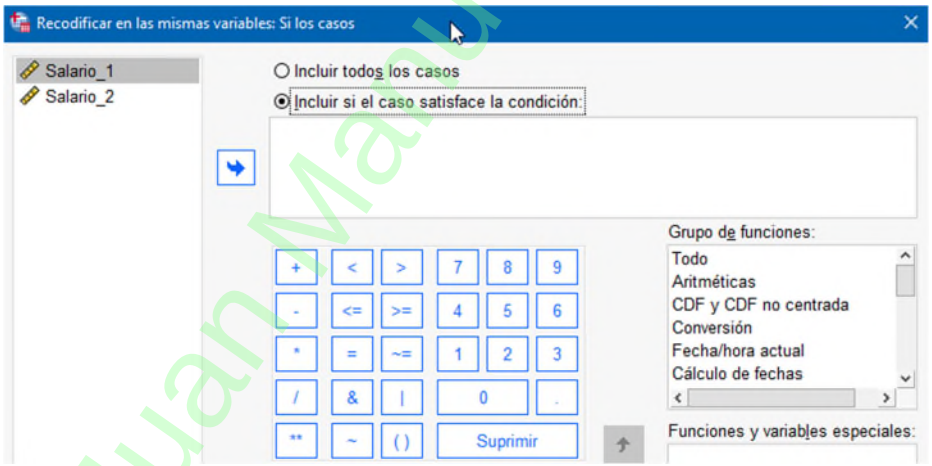
Valor. Valor en el que se va a recodificar uno o más valores antiguos. El tipo de datos (numérico o de cadena) del valor introducido debe coincidir con el tipo de datos del valor antiguo.

Perdido por sistema. Recodifica el valor antiguo especificado como valor perdido por el sistema. El valor perdido por el sistema no se utiliza en los cálculos. Además, los casos con valor perdido por el sistema se excluyen de muchos procedimientos. No está disponible para variables de cadena.

Antiguo->Nuevo. Contiene la lista de especificaciones que se va a utilizar para recodificar la variable o las variables. Puede añadir, cambiar y borrar las especificaciones que desee. La lista se ordena automáticamente basándose en la especificación del valor antiguo y siguiendo este orden: valores únicos, valores perdidos, rangos y todos los demás valores. Si cambia una especificación de recodificación en la lista, el procedimiento volverá a ordenar la lista automáticamente, si fuera necesario, para mantener este orden.



El cuadro de diálogo "Si los casos" (Si la opción), que se utiliza para definir subconjuntos de casos, es el mismo que el descrito para "Calcular variable".



Por ejemplo, vamos a agrupar las variables referidas a salario en las siguientes categorías que representan rangos de salarios:

De 1000 a 2000 --- Categoría 4 (salario bajo)
De 2001 a 3000 --- Categoría 3 (salario medio)
De 3001 a 4000 --- Categoría 2 (salario alto)
De 4001 a 5000 --- Categoría 1 (salario muy alto)

En las figuras anteriores, se ve cómo hay que introducir los valores antiguos y nuevos, de esta forma, los valores de las variables quedan como sigue:

	Salario_1	Salario_2
1	1500,00	1000,00
2	2000,00	3000,00
3	2000,00	4000,00
4	3000,00	2000,00
5	4000,00	5000,00

Antes de recodificar

	Salario_1	Salario_2
1	4,00	4,00
2	4,00	3,00
3	4,00	2,00
4	3,00	4,00
5	2,00	1,00

Después de recodificar

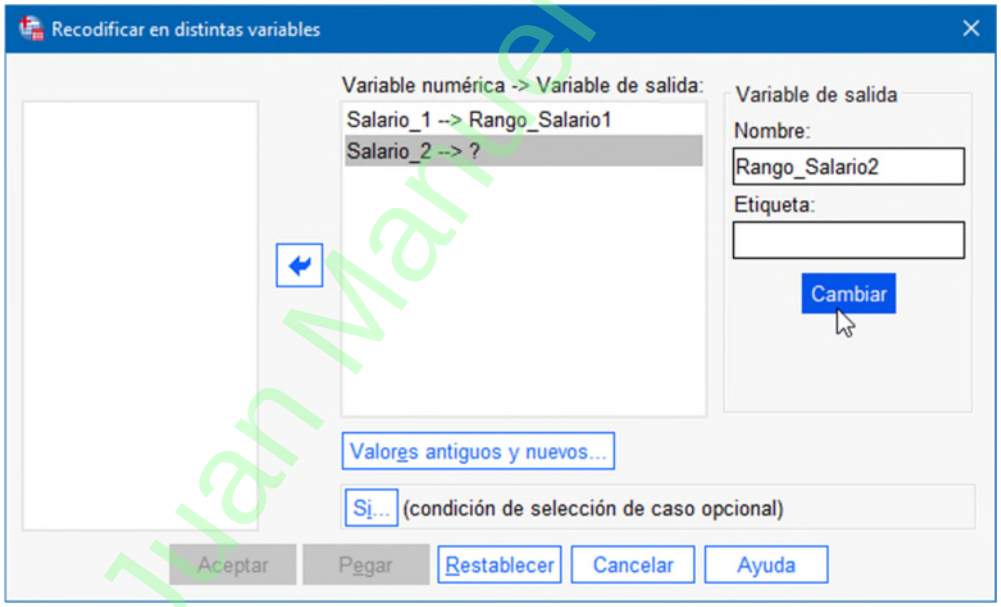
Recodificar en distintas variables

Recodificar en distintas variables reasigna los valores de las variables existentes o agrupa rangos de valores existentes en nuevos valores para una nueva variable.

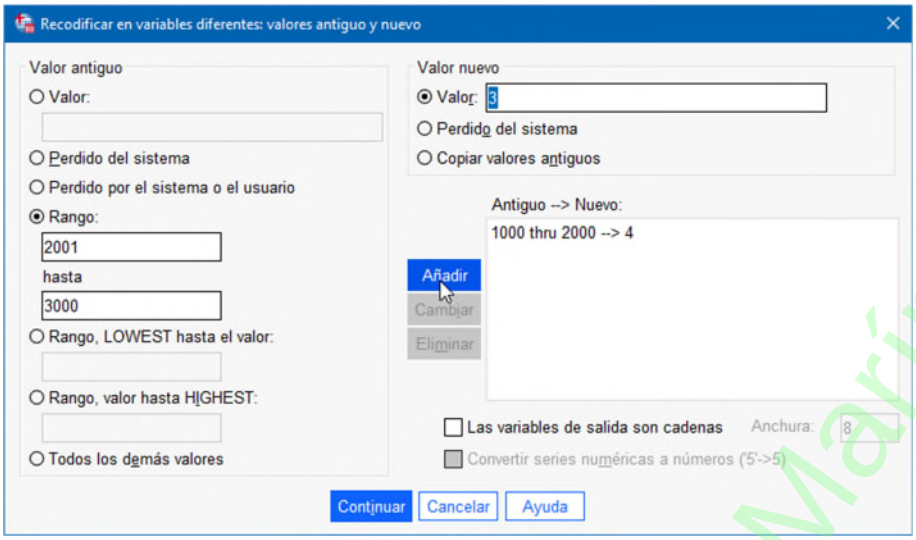
Para recodificar los valores de una variable, elija en los menús: Transformar --- Recodificar --- En distintas variables.

Seleccione las variables que desee recodificar. Si selecciona múltiples variables, todas deberán ser del mismo tipo (numéricas o de cadena).

Introduzca el nombre de la nueva variable de resultado para cada nueva variable y pulse en Cambiar.



Pulse en Valores antiguos y nuevos y especifique cómo deben recodificarse los valores.



Este cuadro de diálogo permite definir los valores que se van a recodificar.

Valor antiguo. Determina el valor o los valores que se van a recodificar. Puede recodificar valores individuales, rangos de valores y valores perdidos. Los rangos y los valores perdidos del sistema no se pueden seleccionar para las variables de cadena, ya que ninguno de los conceptos es aplicable a estas variables. Los valores antiguos deben ser del mismo tipo de datos (numéricos o de cadena) que la variable original. Los rangos incluyen sus puntos finales y los valores definidos como perdidos por el usuario que estén dentro del rango. **Elementos**:

> **Valor.** Valor antiguo individual que se va a recodificar en un valor nuevo. El tipo de datos (numérico o de cadena) del valor introducido debe coincidir con el tipo de datos de las variables que se desea recodificar.
>
> **Perdido por el sistema.** Valores asignados por SPSS cuando los valores de los datos no están definidos de acuerdo con el tipo de formato que se ha especificado, cuando un campo numérico está vacío o cuando no se ha definido un valor generado por un comando de transformación. Los valores numéricos perdidos del sistema se muestran como puntos. Las variables de cadena no pueden tener valores perdidos del sistema, ya que es lícito cualquier carácter en las variables de cadena.
>
> **Perdido por el sistema o por el usuario.** Observaciones que tienen valores que, o bien han sido declarados como perdidos por el usuario o bien son desconocidos y han sido asignados el valor perdido del sistema, lo cual se indica mediante un punto (.).
>
> **Rango.** Un rango inclusivo de valores. No está disponible para variables de cadena. Se incluirán cualesquiera valores perdidos definidos por el usuario que se encuentren dentro del rango.
>
> **Todos los demás valores.** Cualquier valor no incluido en una de las especificaciones de la lista Antiguo-Nuevo. Aparece en la lista Antiguo-Nuevo como ELSE.

Valor nuevo. Es el valor individual en el que se recodifica cada valor o rango de valores antiguo. Los valores nuevos pueden ser numéricos o de cadena.

Valor. Valor en el que se va a recodificar uno o más valores antiguos. El tipo de datos (numérico o de cadena) del valor introducido debe coincidir con el tipo de datos del valor antiguo.

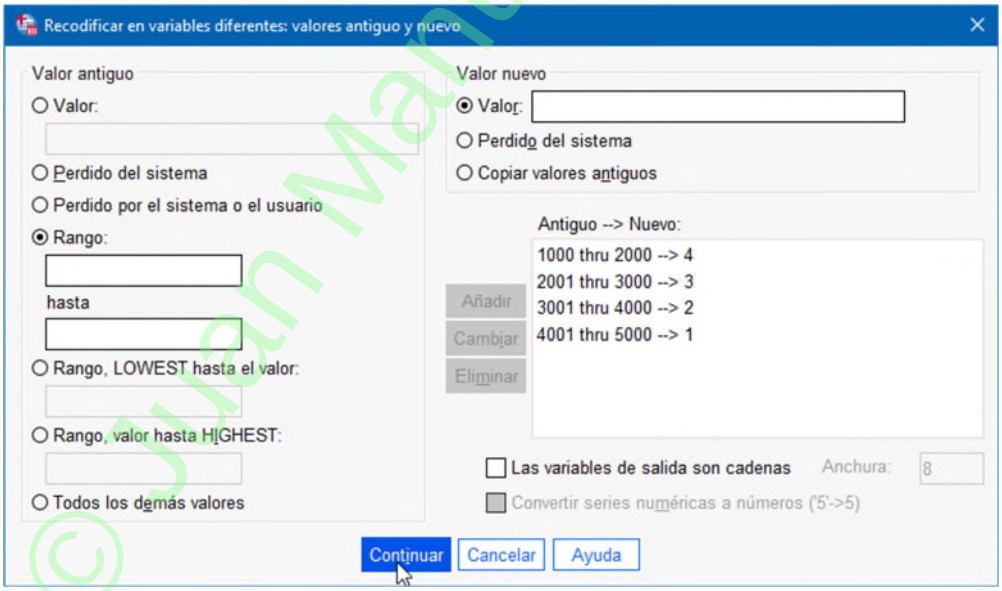
Perdido por el sistema. Recodifica el valor antiguo especificado como valor perdido por el sistema. El valor perdido por el sistema no se utiliza en los cálculos. Además, los casos con valor perdido por el sistema se excluyen de muchos procedimientos. No está disponible para variables de cadena.

Copiar valores antiguos. Retiene el valor antiguo. Si algunos de los valores no requieren la recodificación, utilice esta opción para incluir los valores antiguos. Cualquier valor antiguo no especificado no se incluirá en la nueva variable y a los casos con esos valores se les asignará el valor perdido del sistema para la nueva variable.

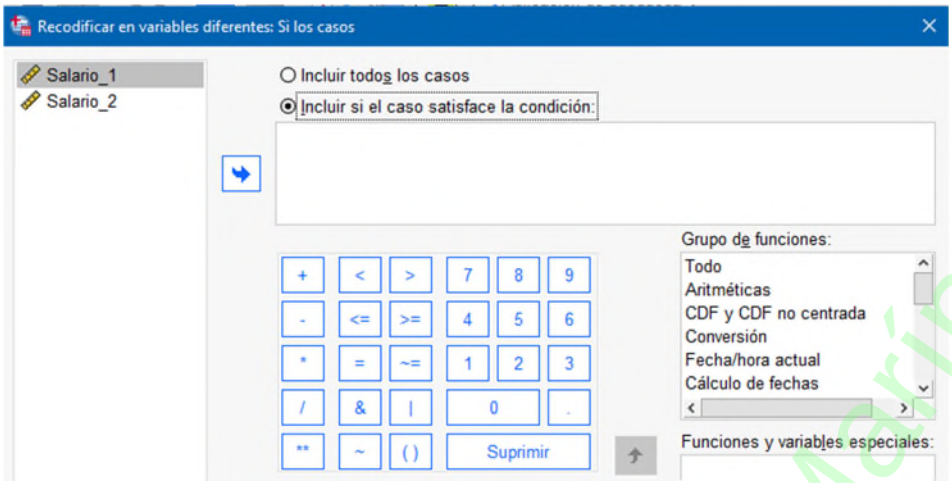
Las variables de resultado son cadenas. Define la nueva variable recodificada como variable de cadena (alfanumérica). La variable antigua puede ser numérica o de cadena.

Convertir cadenas numéricas en números. Convierte los valores de cadena que contienen números en valores numéricos. A las cadenas que contengan cualquier cosa que no sean números y un carácter de signo opcional (+ ó -) se les asignará el valor perdido por el sistema.

Antiguo->Nuevo. Contiene la lista de especificaciones que se va a utilizar para recodificar la variable o las variables. Puede añadir, cambiar y borrar las especificaciones que desee. La lista se ordena automáticamente basándose en la especificación del valor antiguo y siguiendo este orden: valores únicos, valores perdidos, rangos y todos los demás valores. Si cambia una especificación de recodificación en la lista, el procedimiento volverá a ordenar la lista automáticamente, si fuera necesario, para mantener este orden.



El cuadro de diálogo "Si …" (Si los casos), que se utiliza para definir subconjuntos de casos, es el mismo que el descrito para "Calcular variable"



Por ejemplo, vamos a agrupar las variables referidas a salario en las siguientes categorías que representan rangos de salarios:

De 1000 a 2000 --- Categoría 4 (salario bajo)
De 2001 a 3000 --- Categoría 3 (salario medio)
De 3001 a 4000 --- Categoría 2 (salario alto)
De 4001 a 5000 --- Categoría 1 (salario muy alto)

Pero ahora, a diferencia del ejemplo para recodificar en las mismas variables, vamos a crear dos variables nuevas que contendrán la recodificación, llamándolas "Rango_Salario_1" y "Rango_Salario_2".

En las figuras anteriores, se ve cómo hay que introducir los valores antiguos y nuevos, de esta forma, la vista de datos queda como sigue:

	Salario_1	Salario_2
1	1500,00	1000,00
2	2000,00	3000,00
3	2000,00	4000,00
4	3000,00	2000,00
5	4000,00	5000,00

Antes de recodificar

	Salario_1	Salario_2	Rango_Salario1	Rango_Salario2
1	1500,00	1000,00	4,00	4,00
2	2000,00	3000,00	4,00	3,00
3	2000,00	4000,00	4,00	2,00
4	3000,00	2000,00	3,00	4,00
5	4000,00	5000,00	2,00	1,00

Después de recodificar

Recodificación automática

La recodificación automática convierte los valores numéricos y de cadena en valores enteros consecutivos. Si los códigos de la categoría no son secuenciales, las casillas vacías resultantes reducen el rendimiento e incrementan los requisitos de memoria de muchos procedimientos. Además, algunos procedimientos no pueden utilizar variables de cadena y otros requieren valores enteros consecutivos para los niveles de los factores.

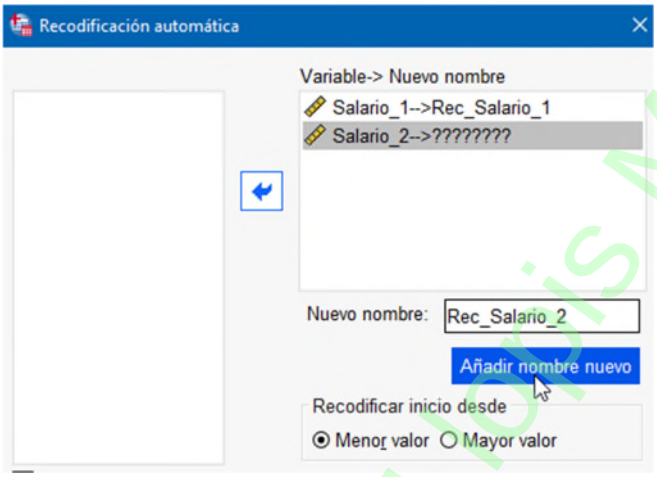
La nueva variable, o variables, creadas por la recodificación automática conservan cualquier variable ya definida y las etiquetas de valor de la variable antigua. Para los valores que no tienen una etiqueta de valor ya definida se utiliza el valor original como etiqueta del valor recodificado. Una tabla muestra los valores antiguos, los nuevos y las etiquetas de valor. Los valores de cadena se recodifican por orden alfabético, con las mayúsculas antes que las minúsculas. Los valores perdidos se recodifican como valores perdidos mayores que cualquier valor no perdido y conservando el orden. Por ejemplo, si la variable original posee 10 valores no

perdidos, el valor perdido mínimo se recodificará como 11, y el valor 11 será un valor perdido para la nueva variable.

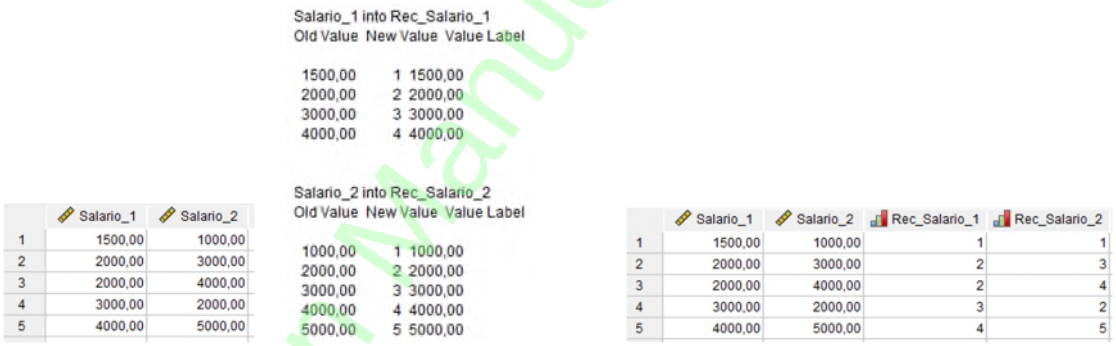
Para recodificar valores numéricos o de cadena en valores enteros consecutivos, elija en los menús: Transformar --- Recodificación automática...

Seleccione la variable o variables que desee recodificar.

Para cada variable seleccionada, introduzca un nombre para la nueva variable y pulse en Nuevo nombre.



Por ejemplo, vamos a recodificar automáticamente las variables "Salario_1" y "Salario_2", empezando por el menor valor, y llamando a las variables resultado "Rec_Salario_1" y "Rec_Salario_2".



	Salario_1	Salario_2
1	1500,00	1000,00
2	2000,00	3000,00
3	2000,00	4000,00
4	3000,00	2000,00
5	4000,00	5000,00

Salario_1 into Rec_Salario_1

Old Value	New Value	Value Label
1500,00	1	1500,00
2000,00	2	2000,00
3000,00	3	3000,00
4000,00	4	4000,00

Salario_2 into Rec_Salario_2

Old Value	New Value	Value Label
1000,00	1	1000,00
2000,00	2	2000,00
3000,00	3	3000,00
4000,00	4	4000,00
5000,00	5	5000,00

	Salario_1	Salario_2	Rec_Salario_1	Rec_Salario_2
1	1500,00	1000,00	1	1
2	2000,00	3000,00	2	3
3	2000,00	4000,00	2	4
4	3000,00	2000,00	3	2
5	4000,00	5000,00	4	5

Antes **Visor de resultados** **Después de recodificar**

TEMA - 4

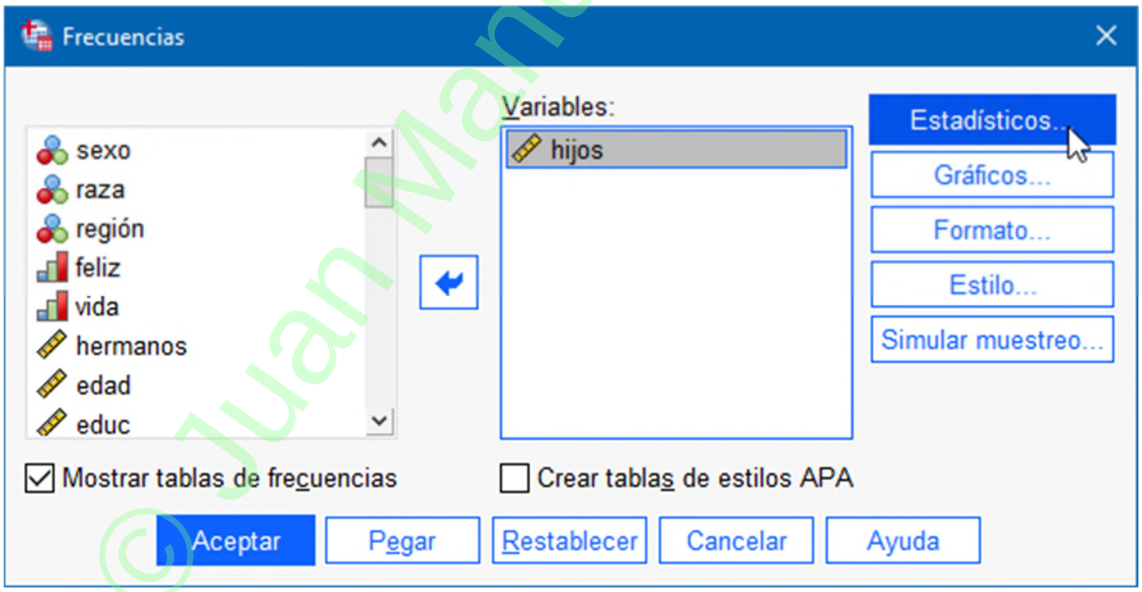
Análisis descriptivos y tablas: Frecuencias, Descriptivos y Tablas cruzadas

1.- Frecuencias

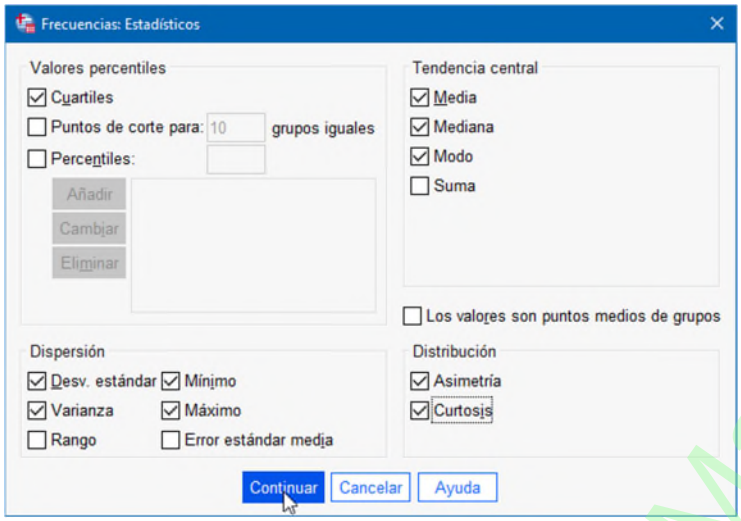
El procedimiento Frecuencias proporciona estadísticos y representaciones gráficas que resultan útiles para describir muchos tipos de variables. Es un buen procedimiento para una inspección inicial de los datos. Para los informes de frecuencias y los gráficos de barras, se pueden organizar los diferentes valores en orden ascendente o descendente u ordenar las categorías por sus frecuencias. Es posible suprimir el informe de frecuencias cuando una variable posee muchos valores diferentes. Se pueden etiquetar los gráficos con las frecuencias (la opción por defecto) o con los porcentajes. **Fichero: encuesta.sav**

Para obtener tablas de frecuencias, elija en los menús: Analizar --- Estadísticos descriptivos --- Frecuencias...

Como **ejemplo**, vamos a calcular las frecuencias para la variable "Número de hijos" (**hijos**) del fichero 'encuesta.sav'.



La siguiente figura muestra los estadísticos que podemos elegir:



Valores percentiles. Los valores de una variable cuantitativa que dividen los datos ordenados en grupos, de forma que un porcentaje de los casos se encuentre por encima y otro porcentaje se encuentre por debajo. Los cuartiles (los percentiles 25, 50 y 75) dividen las observaciones en cuatro grupos de igual tamaño. Si desea un número igual de grupos que no sea cuatro, seleccione Puntos de corte para n grupos iguales (por ejemplo, para calcular los deciles, habría que elegir 10 puntos de corte). También puede especificar percentiles individuales (por ejemplo, el percentil 85, el valor por debajo del cual se encuentran el 85% de las observaciones).

Tendencia central. Los estadísticos que describen la localización de la distribución incluyen: Media, Mediana, Moda y Suma de todos los valores.

> **Media**. Una medida de tendencia central. El promedio aritmético: la suma dividida por el número de casos. El punto en el que se concentra el peso de la distribución de frecuencias. Se la considera el "Centro de Gravedad" de la distribución de frecuencias.
> **Mediana**. Valor por encima y por debajo del cual se encuentran la mitad de los casos; el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos). Se la considera el "Centro Geográfico" de la distribución de frecuencias.
> **Moda**. Valor que ocurre con mayor frecuencia, el que más se repite. Si varios valores comparten la mayor frecuencia de aparición, cada uno de ellas es una moda. El procedimiento de frecuencias devuelve sólo la más pequeña de esas modas múltiples.
> **Suma**. Suma o total de todos los valores, a lo largo de todos los casos que no tengan valores perdidos.

Dispersión. Los estadísticos que miden la cantidad de variación o de dispersión en los datos (el grado en que los valores están próximos entre sí o separados), incluyen: Desviación típica, Varianza, Rango, Mínimo, Máximo y Error típico de la media.

> **Desviación típica**. Es una medida de la dispersión en torno a la media. En una distribución normal, el 68% de los casos se encuentra dentro de una D.T. respecto a la media y el 95% de los casos se encuentra dentro de 2 D.T. respecto a la media. Por ejemplo, si la media de edad es 45, con una desviación típica de 10, el 95% de los casos estaría entre 25 y 65 en una

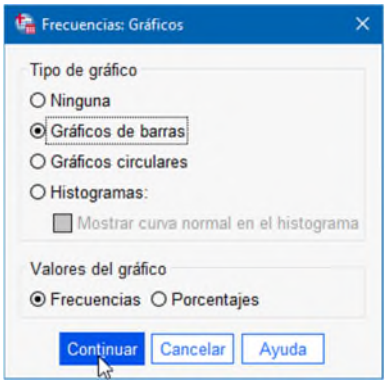
distribución normal. (Realmente, se calcula lo que se llama **Cuasi-Desviación Típica**).
Varianza. Es una medida de dispersión en torno a la media, igual a la suma de las desviaciones al cuadrado respecto a la media, dividida por el número de casos menos 1. La varianza se mide en unidades que son el cuadrado de las de la propia variable. (Realmente, se calcula lo que se llama **Cuasivarianza**).
Amplitud. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.
Mínimo. El menor valor de una variable numérica.
Máximo. El mayor valor de una variable numérica.
E. T. media (Error típico de la media). Es una medida de cuánto puede variar el valor de la media entre varias muestras tomadas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, podremos concluir que los dos valores son distintos si la razón de la diferencia respecto al error típico es menor que -2 o mayor que +2).

Distribución. Asimetría y curtosis son estadísticos que describen la forma y la simetría de la distribución. Estos estadísticos se muestran con sus errores típicos.

Asimetría. Es una medida de la simetría o asimetría de una distribución. La distribución normal es simétrica y tiene un valor de sesgo (asimetría) igual a 0. Una distribución con asimetría positiva (más valores bajos que altos) significativa presenta una cola prolongada hacia la derecha. Una distribución con asimetría negativa (más valores altos que bajos) significativa presenta una cola prolongada hacia la izquierda. De manera aproximada, se considera que un valor de asimetría mayor que dos veces su error típico es indicativo de falta de simetría.
Curtosis. Una medida del grado en que las observaciones se agrupan en torno a un punto central. Una distribución normal tiene curtosis igual a cero (Mesocúrtica). Una distribución con curtosis positiva se dice que es Leptocúrtica y se caracteriza por un centro apuntado y colas engrosadas. Una distribución con curtosis negativa se dice que es Platicúrtica y se caracteriza por un centro plano y colas afinadas.

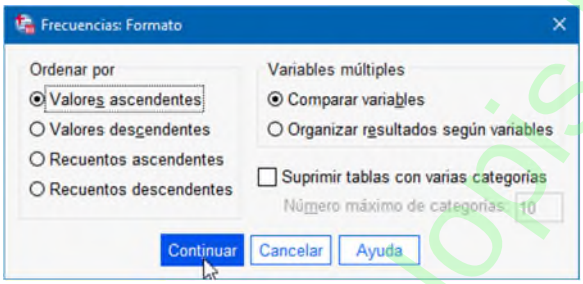
Los valores son puntos medios de grupos. Si los valores de los datos son puntos medios de grupos (por ejemplo, si las edades de todas las personas entre treinta y cuarenta años se codifican como 35), seleccione esta opción para estimar la mediana y los percentiles para los datos originales no agrupados.

En la siguiente figura, vemos los gráficos que podemos elegir:



Tipo de gráfico. Los **gráficos de sectores** muestran la contribución de las partes a un todo. Cada sector de un gráfico de este tipo corresponde a un grupo, definido por una única variable de agrupación. Los **gráficos de barras** muestran la frecuencia de cada valor o categoría distinta como una barra diferente, permitiendo comparar las categorías de forma visual. Los **histogramas** también cuentan con barras, pero se representan a lo largo de una escala de intervalos iguales. La altura de cada barra es el recuento de los valores que están dentro del intervalo para una variable cuantitativa. Los histogramas muestran la forma, el centro y la dispersión de la distribución. Una **curva normal** superpuesta en un histograma ayuda a juzgar si los datos están normalmente distribuidos.
Valores del gráfico. Para los gráficos de barras, puede etiquetar el eje de escala con las **frecuencias** o los **porcentajes**.

La siguiente figura nos muestra el cuadro de diálogo Formato, para indicar qué tipo de ordenación queremos en la tabla de frecuencias:



Ordenar por. La tabla de frecuencias se puede organizar respecto a los valores actuales de los datos o respecto al recuento (frecuencia de aparición) de esos valores, y en orden ascendente o descendente. Sin embargo, si solicita un histograma o percentiles, Frecuencias asumirá que la variable es cuantitativa y mostrará sus valores en orden ascendente.

Múltiples variables. Si desea generar tablas de estadísticos para múltiples variables, podrá mostrar todas las variables en una sola tabla (Comparar variables), o bien mostrar una tabla de estadísticos independiente para cada variable (Organizar resultados según variables).

Suprimir tablas con más de n categorías. Esta opción impide que se muestren tablas que contengan más valores que el número especificado.

En nuestro ejemplo, eligiendo las opciones que aparecen en las figuras anteriores, tendríamos los siguientes resultados en el visor de SPSS:

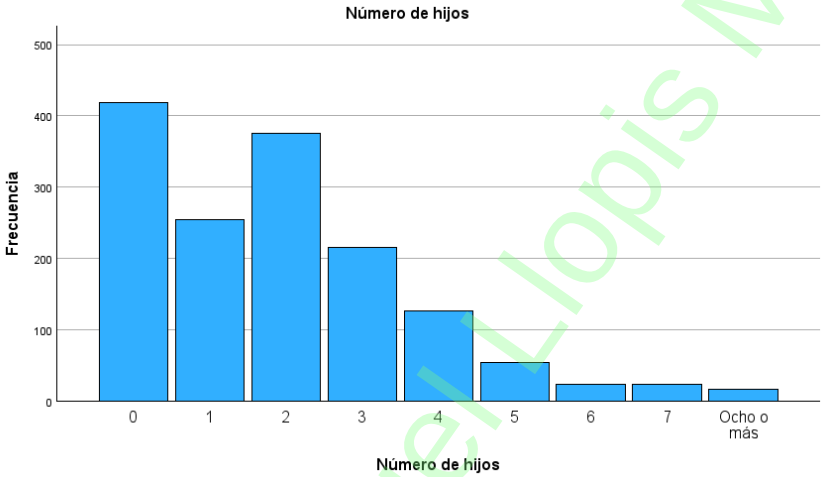
Estadísticos

Número de hijos

N	Válido	1509
	Perdidos	8
Media		1,90
Mediana		2,00
Moda		0
Desv. estándar		1,765
Varianza		3,114
Asimetría		1,034
Error estándar de asimetría		,063
Curtosis		1,060
Error estándar de curtosis		,126
Mínimo		0
Máximo		8
Percentiles	25	,00
	50	2,00
	75	3,00

Número de hijos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
	Total	1509	99,5	100,0	
Perdidos	No contesta	8	,5		
Total		1517	100,0		



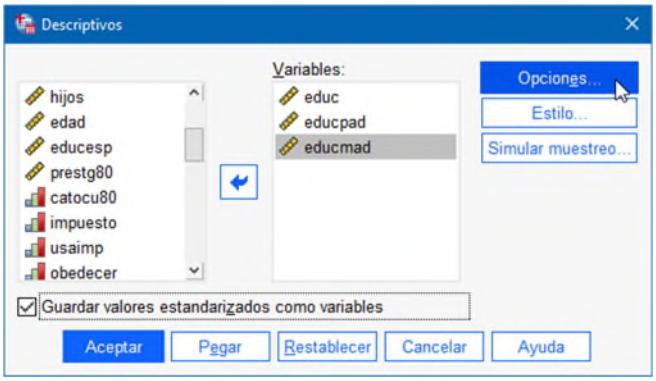
2.- Descriptivos

El procedimiento Descriptivos muestra estadísticos de resumen univariados para varias variables en una única tabla y calcula valores tipificados (puntuaciones z).

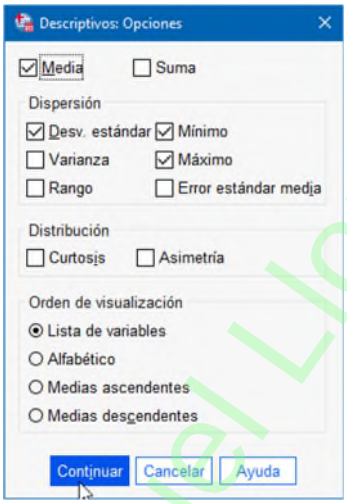
Las variables se pueden ordenar por el tamaño de sus medias (en orden ascendente o descendente), alfabéticamente o por el orden en el que se seleccionen las variables (el valor por defecto).

Cuando se guardan las puntuaciones z, éstas se añaden a los datos del Editor de datos y quedan disponibles para los gráficos, el listado de los datos y los análisis. Cuando las variables se registran en unidades diferentes (por ejemplo, edad y número de hermanos), una transformación en puntuaciones z pondrá las variables en una escala común para poder compararlas visualmente con más facilidad.

Como **ejemplo**, vamos a calcular los descriptivos (guardando los valores tipificados) para las variables "Número de años de escolarización" (**educ**), "Número de años de escolarización del padre" (**educpad**) y "Número de años de escolarización de la madre" (**educmad**) del fichero '**encuesta.sav**'.



Las opciones que aparecen, al pulsar en el botón correspondiente, son las siguientes, ya comentadas en el apartado Frecuencias:



El resultado del visor de SPSS se muestra a continuación:

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. estándar
Número de años de escolarización	1510	0	20	12,88	2,984
Número de años de escolarización del padre	1069	0	20	10,88	4,129
Número de años de escolarización de la madre	1233	0	20	10,79	3,463
N válido (por lista)	973				

En el editor de datos de SPSS, en las columnas finales, aparecerán 3 nuevas variables, que corresponden a las puntuaciones z de las 3 variables utilizadas:

trabajo9	Zeduc	Zeducpad	Zeducmad
0	-,29628	.	,35086
1	2,38467	2,20872	2,08361
2	2,38467	1,23986	,92844
2	2,38467	2,20872	2,66120
0	-,29628	.	.
2	-,96652	-,69787	-1,38189
2	-,96652	-,69787	.
2	1,04419	-1,42452	-1,38189
2	-,96652	-1,18231	-1,67068
0	,37396	-,69787	,35086
0	-1,30163	,27099	.

Estas variables, mantienen el nombre original, con una "Z" al principio.

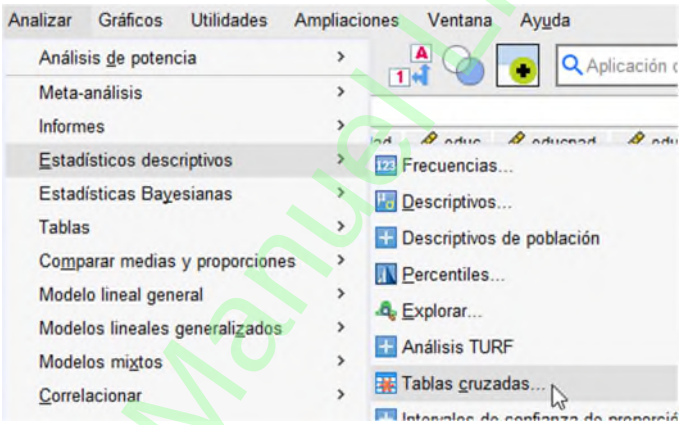
3.- Tablas cruzadas

Las tablas cruzadas nos permiten obtener información fundamentalmente descriptiva acerca de la relación entre los distintos niveles de dos variables nominales u ordinales.

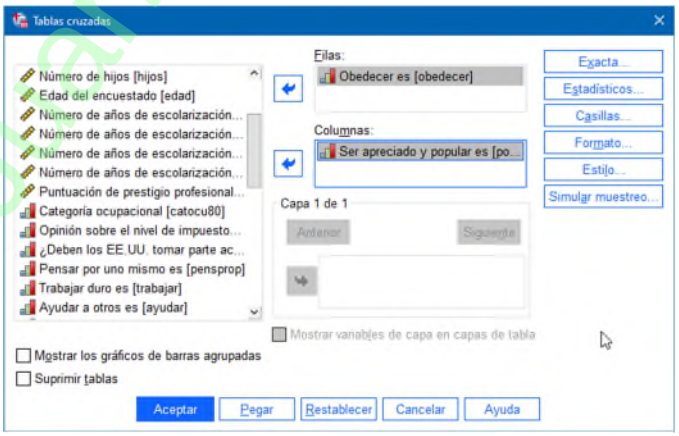
En algunos estadísticos y medidas se asume que hay unas categorías ordenadas (datos ordinales) o unos valores cuantitativos (datos de intervalos o de proporciones), como se explica en el apartado sobre los estadísticos. Otros estadísticos son válidos cuando las variables de la tabla tienen categorías no ordenadas (datos nominales). Para los estadísticos basados en Chi-cuadrado (phi, V de Cramer y coeficiente de contingencia), los datos deben ser una muestra aleatoria de una distribución multinomial.

Las variables ordinales pueden ser códigos numéricos que representen categorías (por ejemplo, 1 = bajo, 2 = medio, 3 = alto) o valores de cadena. Sin embargo, se supone que el orden alfabético de los valores de cadena indica el orden correcto de las categorías. Por ejemplo, en una variable de cadena cuyos valores sean bajo, medio, alto, se interpreta el orden de las categorías como alto, bajo, medio (orden que no es el correcto). Por norma general, se puede indicar que es más fiable utilizar códigos numéricos para representar datos ordinales.

Para acceder al procedimiento tablas cruzadas debemos seguir la ruta habitual *Analizar>Estadísticos descriptivos>tablas cruzadas*:



Apareciendo el menú que podemos observar en la siguiente figura:

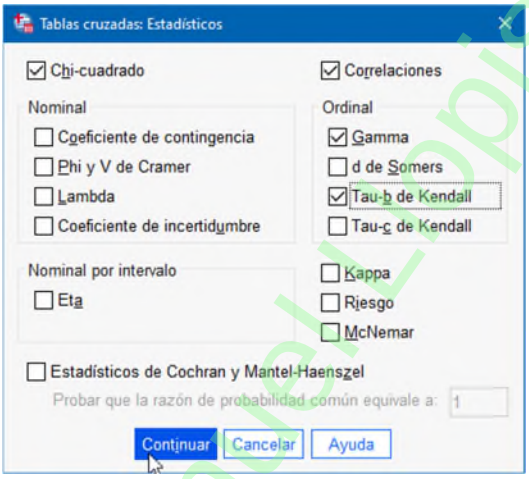


En esta ventana de captura de datos se nos solicita que seleccionemos mediante el procedimiento habitual las variables que deseamos que aparezcan en la fila y

columna de la tabla. Si deseamos que esta tabla se combine con los niveles de otra u otras variables, las introduciremos en la tercera ventana de datos teniendo en cuenta que la salida nos facilitará tan solo tablas de 2 dimensiones, una para cada variable de esta tercera ventana en combinación con las que hemos introducido en fila y columna.
Podemos comprobar cómo en esta ventana tenemos numerosas opciones para completar la información básica que nos proporciona este procedimiento.

Mostrar los gráficos de barras agrupadas. Los gráficos de barras agrupadas ayudan a resumir los datos por grupos de casos. Hay una agrupación de barras por cada valor de la variable especificada en el cuadro Filas. La variable que define las barras dentro de cada agrupación es la variable especificada en el cuadro Columnas. Por cada valor de esta variable hay un conjunto de barras de distinto color o trama. Si especifica más de una variable en Columnas o en Filas, se generará un gráfico de barras agrupadas por cada combinación de dos variables.

Estadísticos.



Correlaciones. Para las tablas en las que tanto las columnas como las filas contienen valores ordenados, en el resultado de Correlaciones hay que fijarse en rho, el coeficiente de correlación de Spearman (sólo datos numéricos). La rho de Spearman es una medida de asociación entre órdenes de rangos. Cuando ambas variables de tabla (factores) son cuantitativas, en el resultado de Correlaciones hay que fijarse en r, el coeficiente de correlación de Pearson, una medida de asociación lineal entre las variables.

Chi-cuadrado. Verifica la hipótesis de nulidad de que las frecuencias observadas en las distintas celdas no difieren de las que serían esperables si no existiera ningún tipo de relación entre las variables. Para las tablas con dos filas y dos columnas, seleccione Chi-cuadrado para calcular la Chi-cuadrado de Pearson, la Chi-cuadrado de la razón de verosimilitud, la prueba exacta de Fisher y la Chi-cuadrado corregido de Yates (corrección por continuidad). Para las tablas 2 × 2, se calcula la prueba exacta de Fisher cuando una tabla (que no resulte de perder columnas o filas en una tabla mayor) presente una casilla con una frecuencia esperada menor que 5. Para las restantes tablas 2 × 2 se calcula la Chi-cuadrado corregida de Yates. Para las tablas con cualquier número de filas y columnas, seleccione Chi-cuadrado para calcular la Chi-cuadrado de Pearson y la Chi-cuadrado de la razón de verosimilitud. Cuando ambas variables de tabla son cuantitativas, Chi-cuadrado da como resultado la prueba de asociación lineal-por-lineal. Además de dicho índice, este procedimiento nos proporciona diversos índices de asociación, la utilización de uno u otro dependerá de la métrica de las variables.

Nominal. Para los datos nominales (sin orden intrínseco, como católico, protestante o judío), puede seleccionar el coeficiente Phi y V de Crámer, el Coeficiente de contingencia, Lambda (lambdas simétricas y asimétricas y tau de Kruskal y Goodman) y el Coeficiente de incertidumbre.

Coeficiente de contingencia. Nos da una idea del grado de asociación entre ambas variables, sus valores oscilaran siempre entre 0 y 1, indicando 1 la máxima asociación. El principal problema de este índice es que el valor máximo que puede alcanzar dependerá del número de filas y columnas, siendo posible alcanzar el valor 1 en el caso hipotético de que el número de filas y columnas fuera infinito.

Phi y V de Cramer. Estas medidas solucionan las limitaciones señaladas por el coeficiente de contingencia, dado que pueden alcanzar el valor máximo para cualquier combinación de filas y columnas. El coeficiente Phi es adecuado en aquellos casos en que la dimensión de la tabla sea 2 x 2, mientras que el coeficiente V lo es en el resto de los casos.

Lambda. Este índice parte de una aproximación diferente a los vistos anteriormente. En lugar de proporcionarnos el grado de asociación entre las variables, nos indica el grado de reducción del error que se produce al utilizar una variable como predictora y otra como criterio. Así por ejemplo si dos variables no están relacionadas, no podemos realizar ninguna predicción sobre las puntuaciones de una de ellas a partir de la otra y, por consiguiente, el error es del 100% y el coeficiente lambda será 0. En el caso contrario, es decir, una relación perfecta entre ambas variables, una es perfectamente predecible a partir de las puntuaciones de la otra, por lo tanto, el error de predicción es 0 y el coeficiente valdrá 1.

Coeficiente de incertidumbre. Medida de asociación que indica la reducción proporcional del error cuando los valores de una variable se emplean para pronosticar los valores de la otra variable. Por ejemplo, un valor de 0,83 indica que el conocimiento de una variable reduce en un 83% el error al pronosticar los valores de la otra variable. SPSS calcula tanto la versión simétrica como la asimétrica del coeficiente de incertidumbre.

Ordinal. Para las tablas en las que tanto las filas como las columnas contienen valores ordenados, seleccione Gamma (orden cero para tablas de doble clasificación y condicional para tablas cuyo factor de clasificación va de 3 a 10), Tau-b de Kendall y Tau-c de Kendall. Para pronosticar las categorías de columna de las categorías de fila, seleccione d de Somers.

Gamma. Medida de asociación simétrica entre dos variables ordinales cuyo valor siempre está comprendido entre menos 1 y 1. Los valores próximos a 1, en valor absoluto, indican una fuerte relación entre las dos variables. Los valores próximos a cero indican que hay poca o ninguna relación entre las dos variables. Para las tablas de doble clasificación, se muestran las gammas de orden cero. Para las tablas de tres o más factores de clasificación, se muestran las gammas condicionales.

d de Somers. Medida de asociación entre dos variables ordinales que toma un valor comprendido entre -1 y 1. Los valores próximos a 1, en valor absoluto, indican una fuerte relación entre las dos variables. Los valores próximos a cero indican que hay poca o ninguna relación entre las dos variables. La d de Somers es una extensión asimétrica de gamma que difiere sólo en la inclusión del número de pares no empatados en la variable independiente. También se calcula una versión simétrica de este estadístico.

Tau-b y Tau-c de Kendall. Coeficientes de correlación no paramétricos que actúan computando el número de inversiones existente entre los rangos de todos los pares de valores para ambas variables. La diferencia entre ambas es la consideración (Tau-b) o no (Tau-c) de los empates. El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la magnitud de la misma, de tal modo que los mayores valores absolutos indican relaciones más fuertes. Los valores posibles van de -1 a 1, pero un valor de -1 o +1 sólo se puede obtener a partir de tablas cuadradas.

Nominal por intervalo. Este caso se produce cuando disponemos de una variable en una escala nominal, como por ejemplo el género y otra en una escala de intervalo o razón, como pueden ser por ejemplo los ingresos mensuales. La variable categórica debe codificarse numéricamente.

Eta. Medida de asociación cuyo valor siempre está comprendido entre 0 y 1. El valor 0 indica que no hay asociación entre las variables de fila y de columna. Los valores cercanos a 1 indican que hay gran relación entre las variables. Eta resulta apropiada para una variable dependiente medida en una escala de intervalo (por ejemplo, ingresos) y una variable independiente con un número limitado de categorías (por ejemplo, sexo). Se calculan dos valores de eta: uno trata la variable de las filas como una variable de intervalo; el otro trata la variable de las columnas como una variable de intervalo, debiendo elegir el resultado que se corresponda con nuestros datos.

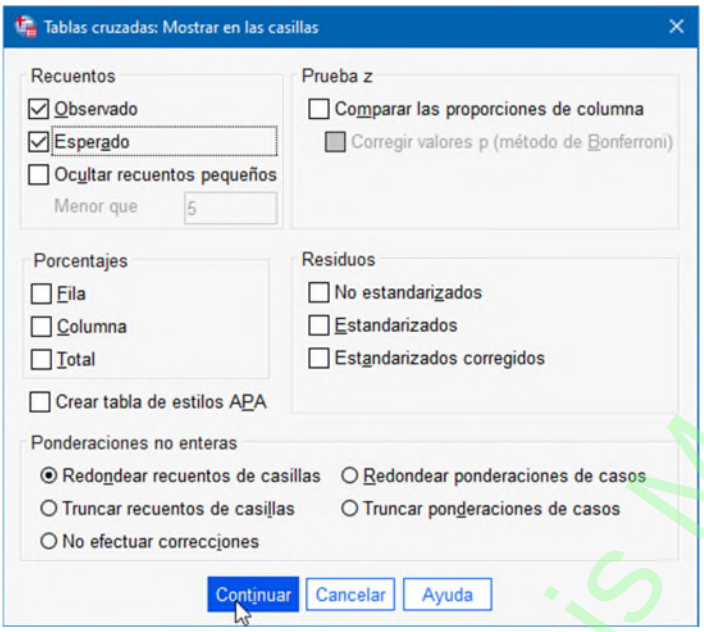
Kappa. Kappa de Cohen mide el acuerdo entre las evaluaciones de dos jueces cuando ambos están valorando el mismo objeto. Un valor igual a 1 indica un acuerdo perfecto. Un valor igual a 0 indica que el acuerdo no es mejor que el que se obtendría por azar. Kappa sólo está disponible para las tablas cuadradas (tablas en las que ambas variables tienen el mismo número de categorías y utilizan los mismos valores de categoría). Este índice es especialmente útil cuando deseemos evaluar el grado de fiabilidad de un sistema de categorías observacional

Riesgo. Para las tablas 2 x 2, medida del grado de asociación entre la presencia de un factor y la ocurrencia de un evento. Si el intervalo de confianza para el estadístico incluye un valor de 1, no se podrá asumir que el factor está asociado con el evento. Cuando la ocurrencia del factor es poco común, se puede utilizar la razón de ventajas como estimación del riesgo relativo.

McNemar. Prueba no paramétrica para dos variables dicotómicas relacionadas. Contrasta los cambios en las respuestas utilizando la distribución de Chi-cuadrado. Es útil para detectar cambios en las respuestas debidas a la intervención experimental en los diseños del tipo "antes-después". Para tablas cuadradas mayores, se utiliza la prueba de simetría de McNemar-Bowker.

Estadísticos de Cochran y de Mantel-Haenszel. Estos estadísticos se pueden utilizar para contrastar la independencia entre una variable dicotómica de factor y una variable dicotómica de respuesta, condicionada por los patrones en las covariables, los cuales vienen definidos por la variable o variables de las capas (variables de control). Hay que tener en cuenta que mientras que otros estadísticos se calculan capa por capa, los estadísticos de Cochran y Mantel-Haenszel se calculan una sola vez para todas las capas.

El procedimiento tablas cruzadas nos permite además obtener información descriptiva sobre las celdas, incluyendo información sobre los residuales. Para ello debemos pulsar el botón "Casillas…" obteniendo la ventana de captura de datos de la figura siguiente:



Así, en este caso hemos pedido que nos muestre las frecuencias observadas para cada celda y las frecuencias esperadas, es decir las que habría si no hubiera ninguna relación entre las variables.

Las opciones que aparecen en la figura anterior son las siguientes:

Frecuencias. El número de casos realmente observados y el número de casos esperados si las variables de fila y columna son independientes entre sí.

Porcentajes. Los porcentajes se pueden sumar a través de las filas o a lo largo de las columnas. También se encuentran disponibles los porcentajes del número total de casos representados en la tabla (una capa).

Residuos. Los residuos brutos no tipificados presentan la diferencia entre los valores observados y los esperados. También se encuentran disponibles los residuos tipificados y tipificados corregidos.

> **No tipificados**. Diferencia entre un valor observado y el valor pronosticado. El valor pronosticado es el número de casos que se esperarían en la casilla si no hubiera relación entre las dos variables. Un residuo positivo indica que hay más casos en la casilla de los que habría en ella si las variables de fila y columna fueran independientes.
>
> **Tipificados**. Es el residuo dividido por una estimación de su desviación típica. Los residuos tipificados, que son conocidos también como los residuos de Pearson o residuos estandarizados, tienen una media de 0 y una desviación típica de 1.
>
> **Tipificados corregidos**. El residuo de una casilla (valor observado menos valor pronosticado) dividido por una estimación de su error típico. El residuo tipificado resultante viene expresado en unidades de desviación típica, por encima o por debajo de la media.

Ponderaciones no enteras. Los recuentos de las casillas suelen ser valores enteros, ya que representan el número de casos de cada casilla. Sin embargo, si el archivo de datos está ponderado en un momento determinado por una variable de ponderación con valores fraccionarios (por ejemplo, 1,25), los recuentos de las casillas pueden

que también sean valores fraccionarios. Puede truncar o redondear estos valores antes o después de calcular los recuentos de las casillas o bien utilizar recuentos de casillas fraccionarios en la presentación de las tablas y los cálculos de los estadísticos.

Redondear frecuencias de casillas. Las ponderaciones de los casos se utilizan tal cual, pero las ponderaciones acumuladas de las casillas se redondean antes de calcular cualquier estadístico.

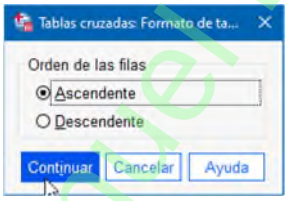
Truncar frecuencias de casillas. Las ponderaciones de los casos se utilizan tal cual, pero las ponderaciones acumuladas de las casillas se truncan antes de calcular cualquier estadístico.

Redondear ponderaciones de casos. Las ponderaciones de los casos se redondean antes de utilizarlas.

Truncar ponderaciones de casos. Las ponderaciones de los casos se truncan antes de utilizarlas.

Sin ajustes. Las ponderaciones de los casos se utilizan sin cambios y se utilizan frecuencias de casillas fraccionarias. Sin embargo, cuando se solicitan estadísticos exactos (disponibles sólo con la opción Pruebas exactas), las ponderaciones acumuladas de las casillas se truncan o se redondean antes de calcular los estadísticos de la prueba exacta.

Pulsando en la opción "**Formato**", puede ordenar las filas en orden ascendente o descendente de los valores de la variable de fila:



Siguiendo las figuras anteriores, puede verse que hemos introducido las variables "obedecer" (etiqueta: Obedecer es) y "popular" (etiqueta: Ser apreciado y popular es), ambas en escala ordinal, en filas y columnas, respectivamente. En "Estadísticos", hemos marcado "Chi-cuadrado", Correlaciones, Gamma y Tau-b de Kendall, dado que tenemos variables ordinales. En "Casillas" hemos marcado el cálculo de las frecuencias observadas y esperadas. El resultado en el visor de SPSS, es el siguiente:

Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Obedecer es * Ser apreciado y popular es	982	64,7%	535	35,3%	1517	100,0%

Tabla cruzada Obedecer es*Ser apreciado y popular es

			Ser apreciado y popular es					
			Lo más importante	Lo 2º más importante	Lo 3º más importante	Lo 4º más importante	Poco importante	Total
Obedecer es	Lo más importante	Recuento	0	12	7	17	159	195
		Recuento esperado	,8	5,4	11,3	36,7	140,8	195,0
	Lo 2º más importante	Recuento	0	0	7	14	102	123
		Recuento esperado	,5	3,4	7,1	23,2	88,8	123,0
	Lo 3º más importante	Recuento	0	5	0	11	126	142
		Recuento esperado	,6	3,9	8,2	26,8	102,5	142,0
	Lo 4º más importante	Recuento	1	5	15	0	322	343
		Recuento esperado	1,4	9,4	19,9	64,6	247,6	343,0
	Poco importante	Recuento	3	5	28	143	0	179
		Recuento esperado	,7	4,9	10,4	33,7	129,2	179,0
Total		Recuento	4	27	57	185	709	982
		Recuento esperado	4,0	27,0	57,0	185,0	709,0	982,0

Tal y como podemos ver, la salida nos ofrece la tabla de cruzada correspondiente y los estadísticos para las celdas que habíamos solicitado.

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	667,504[a]	16	<,001
Razón de verosimilitud	704,769	16	<,001
Asociación lineal por lineal	62,967	1	<,001
N de casos válidos	982		

a. 8 casillas (32,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es ,50.

La salida nos proporciona el valor de Chi-cuadrado y su significación. Podemos comprobar que en este caso y asumiendo un nivel de significación del 5% rechazamos la hipótesis nula de que ambas variables no están relacionadas.

Medidas simétricas

		Valor	Error estándar asintótico[a]	T aproximada[b]	Significación aproximada
Ordinal por ordinal	Tau-b de Kendall	-,365	,030	-11,777	<,001
	Gamma	-,541	,043	-11,777	<,001
	Correlación de Spearman	-,418	,035	-14,403	<,001[c]
Intervalo por intervalo	R de Pearson	-,253	,035	-8,199	<,001[c]
N de casos válidos		982			

a. No se presupone la hipótesis nula.

b. Utilización del error estándar asintótico que presupone la hipótesis nula.

c. Se basa en aproximación normal.

Finalmente, nos ofrece los coeficientes de correlación entre ambas variables. Tal y como podemos comprobar existe una relación moderada entre ambas variables, dado que el coeficiente gamma alcanza un valor de -0,541. De todo ello podemos deducir que ambas variables están relacionadas en el sentido de que las personas que dan más importancia al hecho de ser apreciadas y populares tienden a darle una menor importancia a obedecer órdenes y viceversa.

TEMA - 5

Correlación y Regresión Lineal Simple

1.- Introducción

En este tema se presenta cómo calcular diferentes índices de correlación, así como la forma de modelizar relaciones lineales mediante los procedimientos de regresión simple y múltiple.

Los índices de correlación analizados serán los de Pearson, Spearman, Kendall y el coeficiente de correlación parcial.

Dependiendo de las características de los datos a correlacionar, el coeficiente de correlación que debe aplicarse difiere. Las características principales de cada uno de ellos son las siguientes:

- **Coeficiente de correlación de Pearson**: Es aplicable cuando la métrica de las variables correlacionadas es como mínimo de intervalo.
- **Coeficiente de correlación de Spearman y coeficiente de correlación t de Kendall**: Son aplicables cuando la métrica de las variables es ordinal. El coeficiente de Spearman está especialmente indicado en aquellos casos en que se dé una violación del supuesto de normalidad y, en aquellos casos, en que, aunque la métrica de las variables no sea de intervalo o razón, podemos suponer que la variable con la que trabajamos presenta dicha métrica. P.e. podemos recodificar las puntuaciones de una prueba de inteligencia en tres o cuatro categorías de tal forma que la variable resultante sea ordinal, no obstante, la inteligencia tal y como se ha medido inicialmente mediante un test de CI está en una escala de intervalo. En el resto de los casos es más apropiado utilizar el coeficiente de Kendall.
- **Coeficiente de correlación parcial**: es aplicable cuando se pretende estudiar la relación entre dos variables eliminando el efecto de una tercera variable.

Los procedimientos de regresión, por su parte nos permitirán modelizar la relación existente entre uno (Regresión Lineal Simple) o más predictores (Regresión Lineal Múltiple) con una variable criterio.

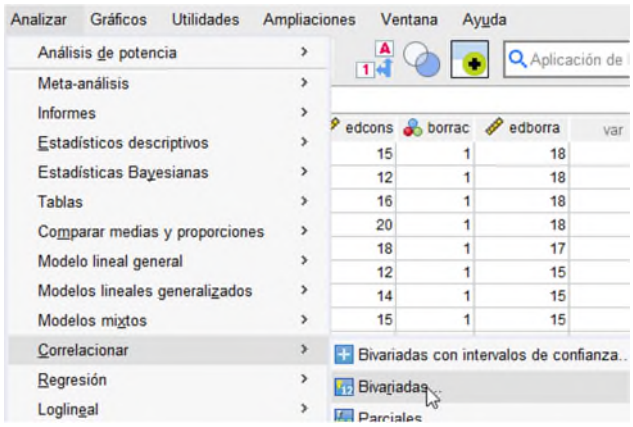
2.- Coeficiente de correlación de Pearson

Este coeficiente es un indicador de la relación lineal existente entre dos variables. El coeficiente de correlación de Pearson es aplicable cuando la métrica de las variables correlacionadas es, como mínimo, de intervalo, y supone que ambas variables se distribuyen en la población de forma normal. No obstante, el coeficiente tan sólo presenta alteraciones destacables en aquellos casos en que se viole de forma considerable dicho supuesto.

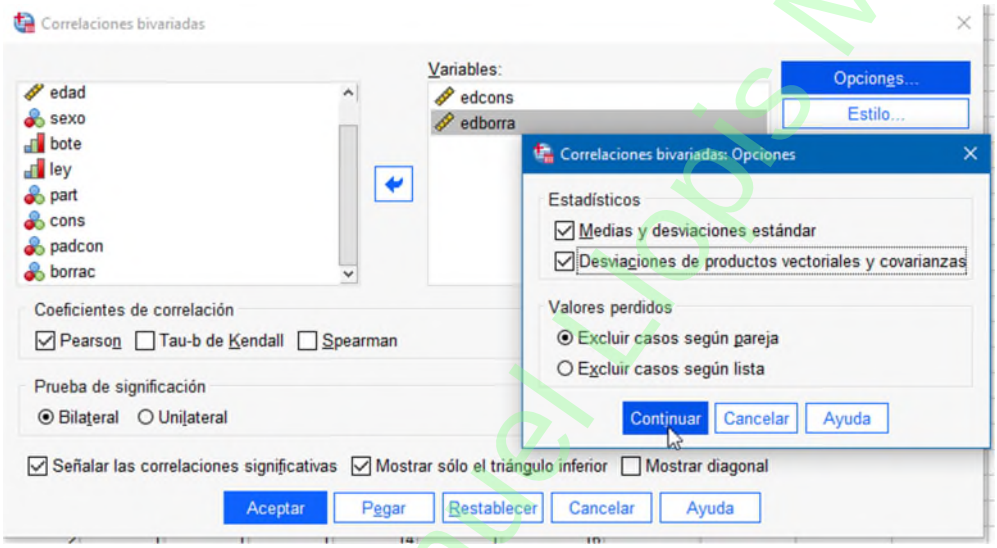
Este coeficiente está acotado, oscila entre -1 y +1, y no posee unidades de medida, lo que hace que su interpretación sea bastante fácil.

Fichero "Botellon.sav"

Para el cálculo de la correlación de Pearson, se procede *Analizar>Correlacionar> Bivariadas*



Aparece la siguiente pantalla.



En esta pantalla marcamos Pearson, y el resto de las alternativas que se indican. En el botón **Opciones**, se nos plantea la posibilidad de seleccionar descriptivos y covarianzas.

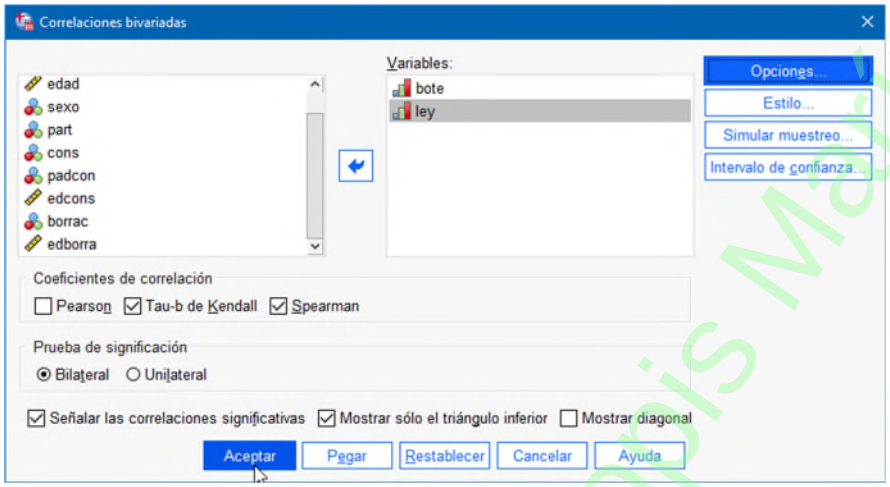
Correlaciones

		6. ¿ A qué edad bebiste alcohol por primera vez?
8. ¿ A qué edad te emborrachaste por primera vez?	Correlación de Pearson	,011
	Sig. (bilateral)	,948
	Suma de cuadrados y productos vectoriales	3,615
	Covarianza	,095
	N	39

La salida nos indica los estadísticos seleccionados, en el ejemplo, vemos que la relación entre las variables consideradas es muy baja, y desde el punto de vista del contraste de hipótesis, nula, porque la significación (Sig.) es mayor que 0,05.

3.- Coeficiente de correlación de Spearman y coeficiente de correlación t de Kendall

Para estudiar en el SPSS los índices de asociación para variables cuasi-cuantitativas (ordinales) se procede como hemos visto en el apartado del Coeficiente de Correlación de Pearson, *Analizar> Correlacionar>Bivariadas*, y aparecen en la pantalla las opciones de Tau-b de Kendall y Spearman.



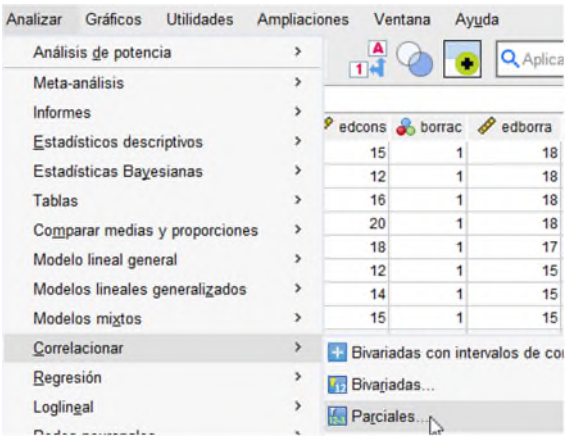
Correlaciones

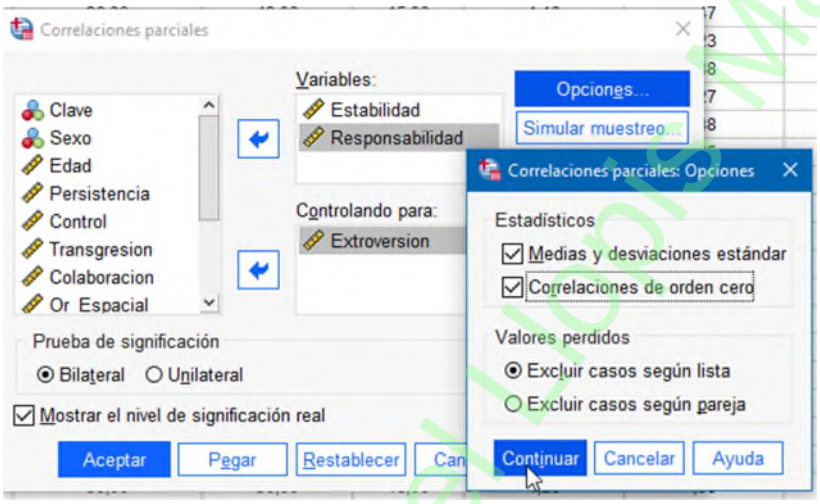
			1. Mi postura ante el fenómeno conocido como el "botellón" es
Tau_b de Kendall	2. Ante la Ley que va a prohibir el consumo de alcohol en la vía pública para todas las personas, mi opinión sobre la Ley es	Coeficiente de correlación	-,040
		Sig. (bilateral)	,769
		N	39
Rho de Spearman	2. Ante la Ley que va a prohibir el consumo de alcohol en la vía pública para todas las personas, mi opinión sobre la Ley es	Coeficiente de correlación	-,034
		Sig. (bilateral)	,835
		N	39

En ambos coeficientes, nos muestra que la relación existente es muy baja y no es significativa (Sig. > 0,05).

4.- Coeficiente de correlación parcial

El coeficiente de correlación parcial se utiliza cuando se pretende estudiar la relación entre dos variables eliminando el efecto de una tercera variable. Esta técnica es útil cuando creemos que la relación existente entre dos variables es debida a una tercera, ya que este coeficiente sería el equivalente a calcular la correlación entre ambas variables manteniendo constante la tercera. Para el cálculo del coeficiente de correlación parcial, se procede como hemos visto en el apartado del Coeficiente de Correlación de Pearson, Analizar>Correlacionar>Parciales, se introducen las variables y se puede obtener estadísticos adicionales en el submenú de opciones. **Fichero "trabajo.sav"**





Correlaciones

Variables de control			Estabilidad Emocional (Neuroticismo)	Responsabilidad (Minuciosidad)	Extroversión
-ninguno-[a]	Estabilidad Emocional (Neuroticismo)	Correlación	1,000	,458	,377
		Significación (bilateral)	.	<,001	<,001
		gl	0	1198	1198
	Responsabilidad (Minuciosidad)	Correlación	,458	1,000	,418
		Significación (bilateral)	<,001	.	<,001
		gl	1198	0	1198
	Extroversión	Correlación	,377	,418	1,000
		Significación (bilateral)	<,001	<,001	.
		gl	1198	1198	0
Extroversión	Estabilidad Emocional (Neuroticismo)	Correlación	1,000	,357	
		Significación (bilateral)	.	<,001	
		gl	0	1197	
	Responsabilidad (Minuciosidad)	Correlación	,357	1,000	
		Significación (bilateral)	<,001	.	
		gl	1197	0	

a. Las casillas contienen correlaciones de orden cero (Pearson).

Es importante señalar que la correlación de Pearson existente entre las variables Estabilidad Emocional y Responsabilidad es de 0,458, mientras que la correlación parcial entre ambas variables se reduce a r=0,357. La conclusión es que la relación entre estabilidad emocional y responsabilidad se reduce por la influencia de la extroversión. En ambos casos la correlación es significativa (Significación ≤ 0,05).

5. Predicción de variables cuantitativas. Concepto de Regresión Lineal Simple y su cálculo con SPSS.

Los términos regresión y predicción se pueden considerar sinónimos en investigación. Se puede utilizar la técnica de la regresión si nos interesa predecir en una población, los valores de una variable "Y" a partir de los valores conocidos de otra variable "X". Dicho objetivo se cubre en dos etapas como sigue:

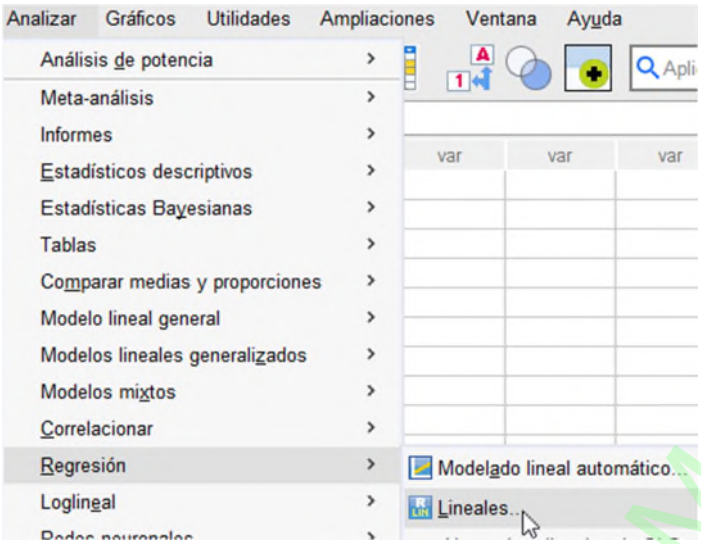
1. Se miden ambas variables en una muestra representativa de la población que nos interesa, y se ajusta la ecuación de la recta que relaciona a ambas variables en dicha muestra.
2. Una vez establecida la ecuación de regresión en la muestra, para predecir la variable Y a partir de X, en cualquier otra muestra procedente de la misma población, sólo se medirá X y se puede estimar el valor de Y sustituyendo X en la ecuación anterior.

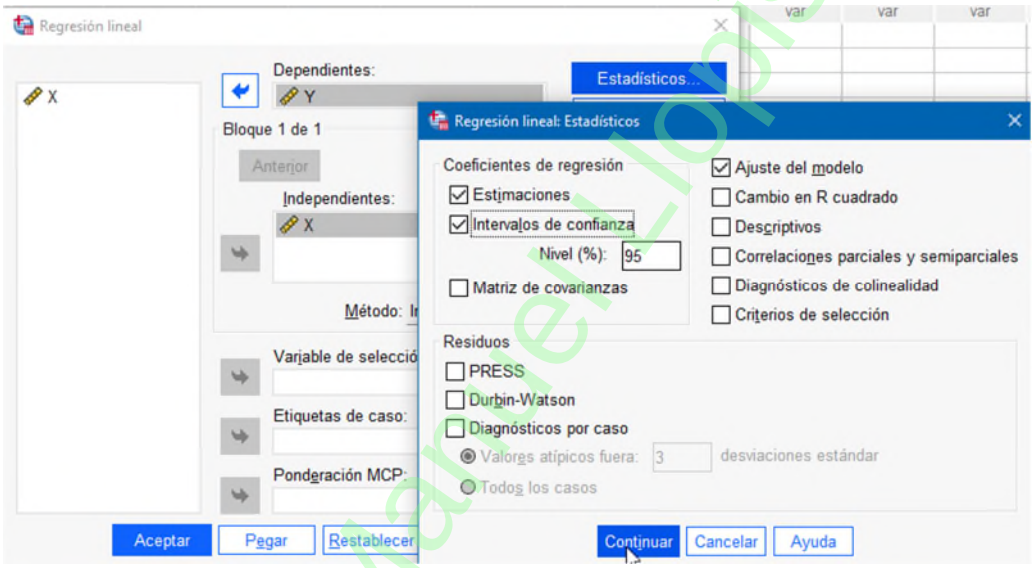
Es importante tener en cuenta que, como la relación entre dos variables no suele ser perfecta, la recta sólo nos dará una predicción aproximada y, por tanto, llevará siempre asociado un término que representa el error de predicción que cometemos con ella, "e".

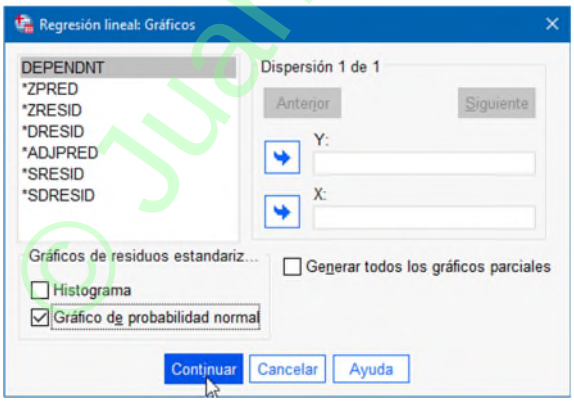
Así partiendo de la ecuación de la recta Y = a + bX, podemos formular el modelo de regresión, que en nuestro caso solo será lineal, $Y = b_0 + b_1X + e$. Donde b_0 es el intercepto, b_1 la pendiente de la recta y "e" es el término de error (error típico de la estimación), es decir, la diferencia entre la puntuación predicha por el modelo y la observada.

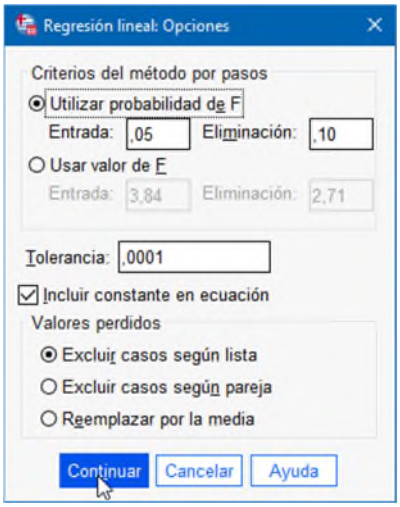
Un indicador del porcentaje de explicación de la variable Y a partir de la variable X lo ofrece el coeficiente de determinación o cuadrado del coeficiente de correlación de Pearson. Veamos un ejemplo de dos variables X (Tasa Cardíaca) e Y (autopercepción de síntomas del síndrome de abstinencia). La variable Y posee una escala de 1-10, siendo el 1 muy poca autopercepción. A continuación, se exponen los diferentes menús y submenús para realizar una regresión lineal simple mediante *Analizar>Regresión>Lineales*, y vamos solicitando las diferentes opciones en los distintos botones de Estadísticos, Gráficos, Guardar y Opciones. Muchas de las alternativas que se ofrecen en este menú son para la evaluación y diagnóstico de la regresión múltiple, que no vamos a ver. Así que estudiaremos la salida con las opciones que hemos marcado en cada uno de los botones que aparecen en la siguiente figura, así como los elementos de la salida que SPSS realiza por defecto. Los datos utilizados son los siguientes: **Fichero: "TasaCardiaca.sav"**.

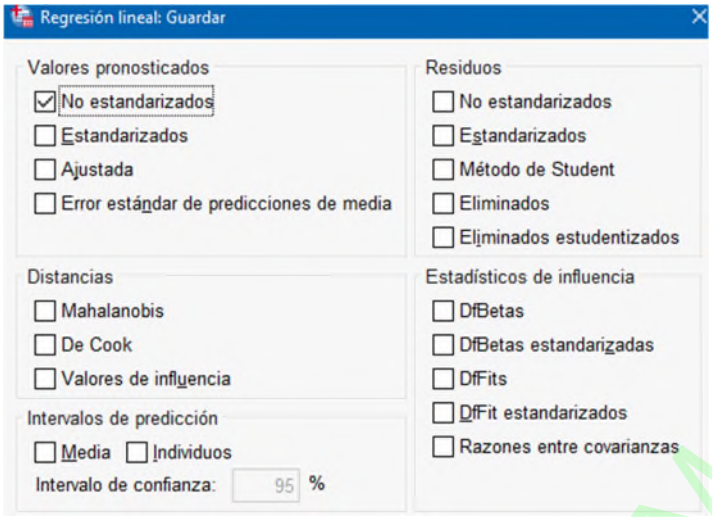
X	70	70	85	88	110	135	140	140	145
Y	3	4	2	4	7	8	8	7	9











La salida correspondiente a las opciones marcadas, en primer lugar, nos señala el porcentaje de varianza explicada, que es el coeficiente de determinación (Coeficiente de correlación de Pearson al cuadrado), así como un ANOVA de la Regresión.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,917[a]	,842	,819	1,080

a. Predictores: (Constante), Tasa cardíaca

b. Variable dependiente: Autopercepción de síntomas del síndrome de abstinencia

R cuadrado, es considerado el índice de bondad de ajuste del modelo de regresión.

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	43,388	1	43,388	37,187	<,001[b]
	Residuo	8,167	7	1,167		
	Total	51,556	8			

a. Variable dependiente: Autopercepción de síntomas del síndrome de abstinencia

b. Predictores: (Constante), Tasa cardíaca

A continuación, nos indica las diferentes estimaciones de los parámetros de la ecuación de regresión de tal manera que Y' = -2,289 + 0,074X + 1,080.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B	
		B	Desv. Error	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	-2,289	1,371		-1,670	,139	-5,531	,953
	Tasa cardíaca	,074	,012	,917	6,098	<,001	,045	,102

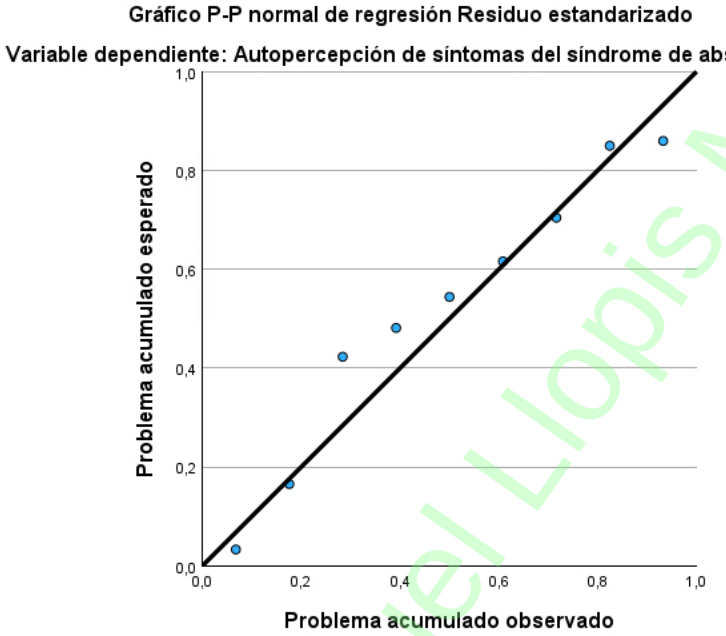
a. Variable dependiente: Autopercepción de síntomas del síndrome de abstinencia

Posteriormente, se evalúan los residuos mediante índices numéricos y gráficos. Se tienen en cuentan los residuos tipificados a partir de +3 y -3, que no se dan en este caso.

Estadísticas de residuos[a]

	Mínimo	Máximo	Media	Desv. estándar	N
Valor pronosticado	2,88	8,42	5,78	2,329	9
Residuo	-1,989	1,165	,000	1,010	9
Desv. Valor pronosticado	-1,244	1,135	,000	1,000	9
Desv. Residuo	-1,841	1,078	,000	,935	9

a. Variable dependiente: Autopercepción de síntomas del síndrome de abstinencia



Los datos se ajustan al modelo lineal

Como en "guardar" habíamos marcado la opción de "**valores pronosticados – no estandarizados**", en el fichero de datos nos aparece una columna con los valores que tendría la variable dependiente (Y), al aplicar la ecuación de regresión sobre la variable independiente (X), esto es, los valores pronosticados de Y.

	X	Y	PRE_1
1	70	3	2,88085
2	70	4	2,88085
3	85	2	3,98874
4	88	4	4,21032
5	110	7	5,83522
6	135	8	7,68171
7	140	8	8,05100
8	140	7	8,05100
9	145	9	8,42030

6.- Interpretación del Coeficiente de Correlación de Pearson y del Coeficiente de Regresión

El coeficiente de correlación de Pearson, y en general todos los coeficientes utilizados, no significan "causalidad". Es decir, nos indican asociación, en su mayoría lineal entre dos variables, pero no implica que una variable sea la causa de la otra. En el contexto de la regresión simple, el coeficiente de regresión estandarizado (beta) coincide con el coeficiente de correlación de Pearson. El cuadrado de este valor es el coeficiente de determinación y se interpreta como la proporción de varianza asociada de una variable sobre la otra.

La magnitud del coeficiente de correlación de Pearson se interpreta desde los estudios realizados anteriormente, y dependiendo de las variables con las que se está trabajando. No obstante, a título meramente orientativo puede servir la siguiente tabla:

Coeficiente de Correlación	Descripción
$0 < r_{xy} \leq 0,2$	Muy Baja
$0,2 < r_{xy} \leq 0,4$	Baja
$0,4 < r_{xy} \leq 0,7$	Moderada o media
$0,7 < r_{xy} \leq 0,9$	Alta
$0,9 < r_{xy} < 1$	Muy alta

Esta tabla sirve tanto para valores positivos como negativos. En caso de ser igual a +1 la correlación será Positiva Perfecta. Si el coeficiente es igual a -1 tendremos una correlación Negativa Perfecta.

TEMA - 6

Inferencia Estadística

1.- Introducción

1.1. Características fundamentales de los diseños básicos con dos muestras (grupos) independientes.

Los diseños de investigación con dos muestras o dos grupos independientes son las estructuras de investigación más básicas, tanto en el ámbito experimental como no experimental. En el contexto experimental, quedan caracterizados por la existencia de único factor o variable independiente directamente manipulado por el investigador y una variable dependiente o variable respuesta. La variable independiente toma dos niveles, a los cuales son asignados los sujetos participantes en la investigación. Esta asignación debe ser aleatoria para asegurar la equivalencia inicial de los grupos, formándose de esta manera dos grupos o muestras independientes que, posteriormente a la aplicación del tratamiento experimental, serán comparados habitualmente a través de la puntuación media de cada grupo en la VD.

En el contexto no experimental, la estructura básica del diseño es la misma, con la diferencia de que, en este caso, la variable independiente no puede ser directamente manipulada ni los grupos pueden ser asignados aleatoriamente a los dos niveles de la variable independiente. Por tanto, en los diseños no experimentales de dos grupos o dos muestras independientes, se trabaja habitualmente con dos grupos ya formados o existentes, cuyas medias o medidas similares en una variable respuesta (VD) interesa comparar.

1.2. Características fundamentales de los diseños básicos con dos muestras relacionadas (diseños con medidas repetidas)

Las estructuras de investigación en las que tenemos "dos muestras relacionadas" corresponden a los diseños básicos con medidas repetidas y, como en el caso anterior, estas estructuras de investigación pueden tener tanto un carácter experimental como no experimental.

En el contexto experimental, los diseños básicos con medidas repetidas, denominados también "intrasujetos", quedan caracterizados, al igual que los diseños de dos grupos, por la existencia de un único factor o variable independiente directamente manipulado por el investigador, que toma dos niveles, y un variable dependiente o variable respuesta. Pero a diferencia de ellos, cada uno de los sujetos participantes en la investigación es sometido a los dos niveles de la variable independiente, generando cada sujeto dos puntuaciones en lugar de una.
En estudios no experimentales, la estructura básica del diseño es la misma, con la diferencia de que, en este caso, la variable independiente no puede ser directamente manipulada, y lo que interesa habitualmente es medir al mismo grupo de sujetos en dos ocasiones distintas en una misma variable respuesta o VD, para posteriormente comparar ambas mediciones.

2.- Introducción. Organización de los datos

Como hemos visto anteriormente, una de las estrategias básicas de análisis acopladas al diseño es la comparación de los promedios o índices similares en los grupos objeto de nuestro estudio. El paquete estadístico SPSS proporciona pruebas t para la diferencia de dos medias independientes, dependientes y para una única

media. El estadístico adecuado en cada caso es diferente y, por tanto, el procedimiento a seguir y también la forma de introducir los datos es distinta.

a. Diferencia de dos medias independientes: En este caso, lo que pretendemos es saber si una variable tiene una media diferente en dos grupos o muestras distintas. Por ejemplo, deseamos saber si la clase A, que recibe su enseñanza por la mañana tiene por término medio más inteligencia (Cociente Intelectual) que el grupo que recibe clases por la tarde o clase B. Así, en este caso estamos contrastando dos grupos diferentes de sujetos y por ello se denominan independientes. Los datos se organizan siguiendo el esquema general habitual del SPSS de un caso para cada sujeto y una fila para cada variable. En nuestro caso los datos quedarían así:

Sujeto	Grupo	C.I.
1	A	101
2	B	103
3	A	98
4	A	105
5	B	99

Es decir, una variable sería la variable de agrupamiento o VI (en nuestro caso la variable Grupo) y la segunda variable sería la variable sobre la que pretendemos saber si hay diferencias o VD (en nuestro caso la variable C.I.).

b. Diferencia de dos medias dependientes o relacionadas: En este caso, pretendemos contrastar las medias de una variable que se ha medido dos veces en los mismos sujetos. Por ejemplo, queremos saber si, por término medio, la tasa cardíaca es mayor en un grupo de sujetos por la mañana (TCM) o a última hora del día (TCT). Para ello elegimos una muestra de sujetos y les medimos esa variable en dos ocasiones. Así, en este caso, la muestra es la misma solo que hemos medido en dos ocasiones, por ello también se denomina para medias relacionadas o pareadas o de medidas repetidas. En este caso, los datos se introducen de forma distinta, pero siguiendo el esquema general, una fila para cada caso o sujeto y una columna para cada variable.

Sujeto	TCM	TCT
1	58	65
2	72	72
3	64	73
4	68	80
5	67	63

En este caso, no hay variable para hacer grupos, pero la variable tasa cardíaca, al haberse medido en dos ocasiones distintas, ocupa dos columnas distintas, y contrastaremos la media de una columna con la otra.

c. Diferencia de una sola media: En este caso, lo que se pretende es contrastar si una variable de un grupo de sujetos medido en una sola ocasión, es decir, una media, difiere significativamente o no de un determinado valor prefijado. Por ejemplo, desearíamos saber si un grupo de sujetos difiere significativamente de la edad de 22 años. Los datos pues, se organizan de la siguiente forma:

Sujeto	Edad
1	25
2	30
3	18
4	22
5	24

En esta situación, sólo hay una columna para la variable que deseamos contrastar.

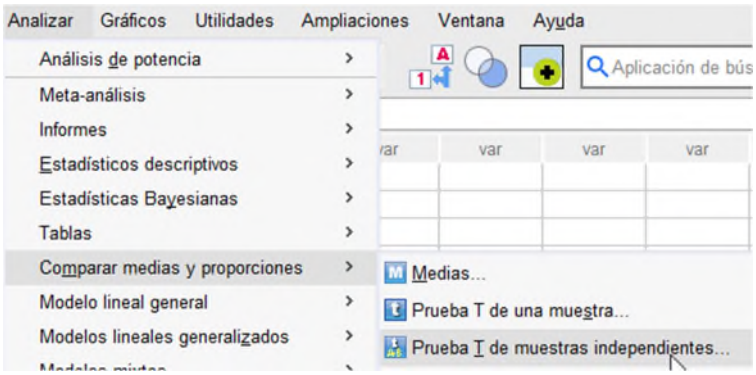
Para ejecutar cualquiera de las tres opciones apuntadas, el procedimiento a elegir será dentro del menú Analizar, y dentro de este, el procedimiento Comparar medias y proporciones. Veamos más detenidamente las opciones y posibilidades de cada procedimiento.

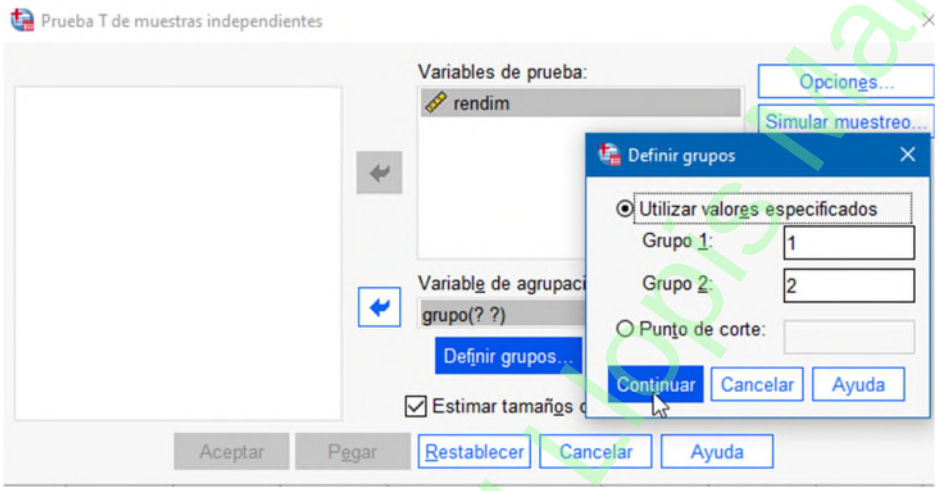
3. Diferencia de dos medias independientes

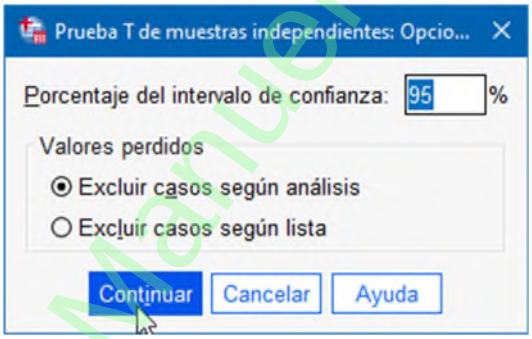
Este tipo de contraste se puede aplicar a una investigación experimental con una estructura de Diseño unifactorial aleatorios de dos grupos. Donde un investigador desea comprobar si la presencia de ruido moderado afecta y en qué modo, a la ejecución de una tarea de resolución de problemas de razonamiento. Para ello, se selecciona una muestra de 18 sujetos, que son divididos al azar en dos grupos. A su vez, se adjudica al azar las condiciones de tratamiento: los que son sometidos a un nivel moderado de ruido (similar a la sintonización de una radio) mientras realizan la tarea (grupo 1) y los que realizan en ausencia de ruido (grupo 2). El rendimiento en la tarea proporciona la medición de la variable dependiente. El nivel de significación considerado es alfa = 0,05. **Fichero "2mind.sav"**.

Una vez introducidos los datos (en dos columnas una para la variable grupo y otra para la variable rendimiento y en 18 filas, una para cada sujeto) seleccionaremos el procedimiento *Analizar>Comparar medias y proporciones>Prueba T de muestras independientes*. Tras seguir estos pasos, nos aparece una ventana. En esa ventana debemos seleccionar las variables para el análisis. En el cuadro **Contrastar variables** debemos incluir la variable que queremos contrastar (en nuestro caso Rendimiento) y en el cuadro **Variable de agrupación** se incluye la variable con la que se definen los grupos (en nuestro caso la variable que hemos denominado grupo).

Cuando se selecciona esta variable de agrupación, obligatoriamente se debe pulsar el botón **Definir grupos**. Los grupos que hemos definido para el análisis y que deben coincidir con los valores incluidos en el Editor de datos cuando hemos definido la variable. Además, en esta ventana aparece el botón **Opciones**. Permite seleccionar el tanto por ciento para el intervalo de confianza con el que queremos contrastar nuestra hipótesis y excluir casos según el análisis o según lista. Una vez seleccionadas correctamente las opciones anteriores se proporciona la salida.







Estadísticas de grupo

	grupo	N	Media	Desv. estándar	Media de error estándar
rendim	Ruido moderado	9	23,78	2,048	,683
	Ausencia de ruido	9	28,67	2,000	,667

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias					
						Significación			
		F	Sig.	t	gl	P de un factor	P de dos factores	Diferencia de medias	Diferencia de error estándar
rendim	Se asumen varianzas iguales	,007	,934	-5,124	16	<,001	<,001	-4,889	,954
	No se asumen varianzas iguales			-5,124	15,991	<,001	<,001	-4,889	,954

Como puede observarse la salida proporciona los estadísticos de los dos grupos estudiados y dos test, el de Levene, que permite contrastar la igualdad o desigualdad de las varianzas, es decir el supuesto de homogeneidad de varianzas entre los grupos

(homocedasticidad) y la prueba t de diferencia de medias en la que se indican el valor del estadístico de contraste (t), los grados de libertad del contraste (gl), la probabilidad para el contraste bilateral (Sig.), así como el error típico de medida para la diferencia de medias y el intervalo de confianza para la diferencia (que nos permite ver los límites entre los que se encuentra el verdadero valor del parámetro de la diferencia de medias que no aparece en la captura por problemas de tamaño). En este caso cabría concluir que existen diferencias estadísticamente significativas entre los grupos (Sig. < 0,001), pero lo más importante es cuantificar su magnitud (tamaño del efecto). En las últimas versiones, SPSS lo calcula por defecto.

Tamaños de efecto de muestras independientes

		Standardizer[a]	Estimación de puntos	Intervalo de confianza al 95% Inferior	Superior
rendim	d de Cohen	2,024	-2,415	-3,636	-1,153
	corrección de Hedges	2,126	-2,300	-3,463	-1,098
	delta de Glass	2,000	-2,444	-3,913	-,922

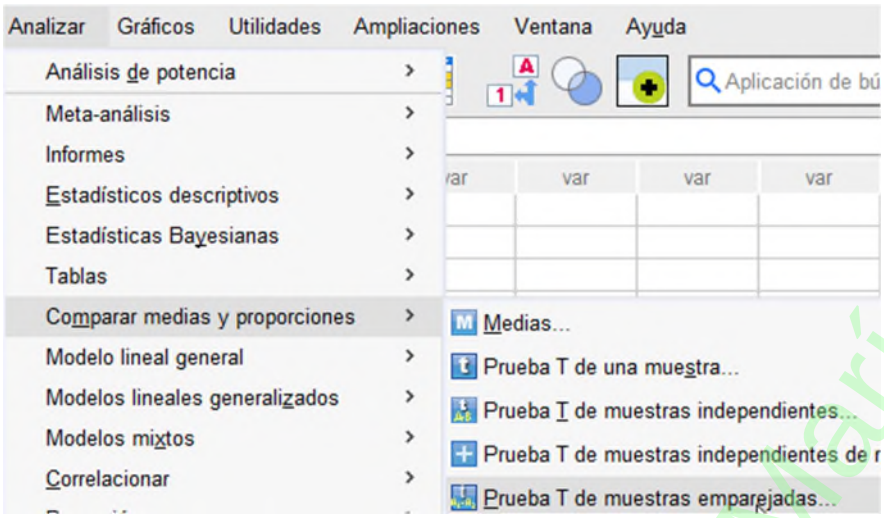
a. El denominador utilizado en la estimación de tamanos del efecto.
d de Cohen utiliza la desviación estándar combinada.
La corrección de Hedges utiliza la desviación estándar combinada, más un factor de corrección.
Delta de Glass utiliza la desviación estándar de la muestra del grupo de control (es decir, el segundo).

El índice d de Cohen (Estimación de puntos, en positivo: 2,415) es un estimador del tamaño del efecto. Como guía orientativa del tamaño del efecto obtenido mediante índices de asociación, se puede seguir la siguiente tabla:

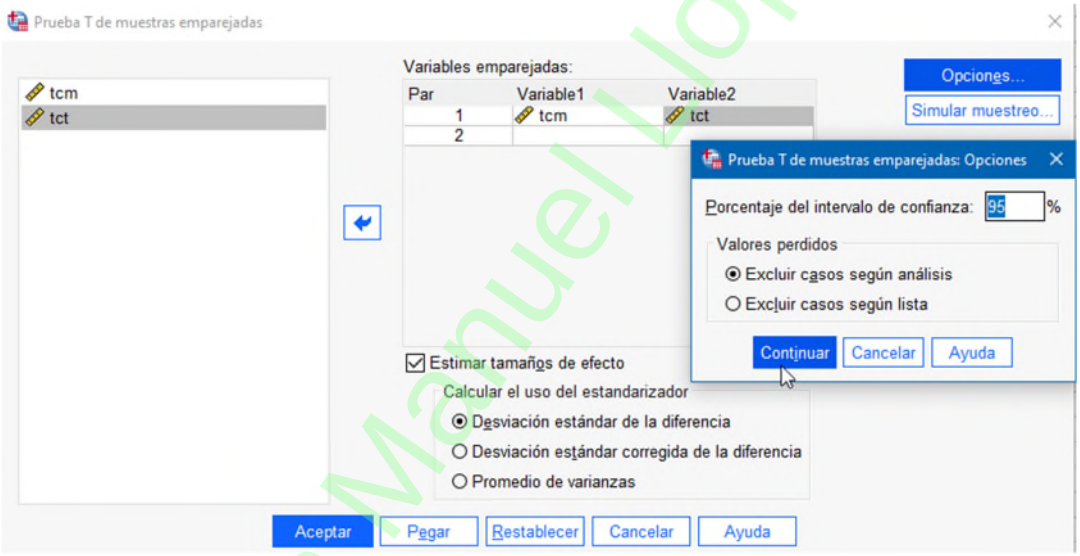
Tamaño del efecto	Intrascendente	Pequeño	Moderado	Alto	Muy alto	Casi perfecto	Perfecto
d	0.0	0.2	0.6	1.2	2.0	4.0	infinite

4.- Diferencia de dos medias dependientes o relacionadas

En este caso, los pasos a seguir son similares al de medias independientes, aunque, como ya hemos aclarado previamente, los datos se introducen de forma distinta. Vamos a aplicar esta prueba para analizar un Diseño Experimental Unifactorial de medidas repetidas, con dos medidas. Supongamos que tenemos una muestra de 20 sujetos deportistas a los que les hemos medido su tasa cardíaca por la mañana antes de levantarse de la cama y horas más tarde después de haber realizado un entrenamiento. Queremos saber si la fatiga después de entrenar influye en la tasa cardíaca. Para ello introducimos los datos obtenidos y seleccionamos sucesivamente en el menú *Analizar>Comparar medias y proporciones>Prueba T de muestras emparejadas*. **Fichero "2mrel.sav"**.



En esta ventana, a diferencia de las anteriores y dado que necesitamos dos variables para contrastar, es necesario seleccionar dos variables antes de incluirlas en el cuadro de Variables relacionadas. El botón opciones permite modificar el intervalo de confianza y por si ha perdido datos de una o más variables, puede indicar el procedimiento de casos que desea excluir.



La salida que proporciona el programa nos da la media para cada variable, la correlación entre ambas y la prueba t que indica si existen o no diferencias estadísticamente significativas.

Estadísticas de muestras emparejadas

		Media	N	Desv. estándar	Media de error estándar
Par 1	tasa cardíaca mañana	64,90	20	11,706	2,618
	tasa cardíaca tarde	68,95	20	13,189	2,949

Correlaciones de muestras emparejadas

		N	Correlación	Significación P de un factor	Significación P de dos factores
Par 1	tasa cardíaca mañana & tasa cardíaca tarde	20	,978	<,001	<,001

Prueba de muestras emparejadas

		Diferencias emparejadas					t	gl	Significación P de un factor	Significación P de dos factores
		Media	Desv. estándar	Media de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior				
Par 1	tasa cardíaca mañana - tasa cardíaca tarde	-4,050	2,982	,667	-5,446	-2,654	-6,074	19	<,001	<,001

En este caso habría que concluir que existen diferencias estadísticamente significativas entre los grupos, pero lo más importante es cuantificar su magnitud (tamaño del efecto), el tamaño del efecto aparece en la siguiente tabla de resultados de SPSS:

Tamaños de efecto de muestras emparejadas

			Standardizer[a]	Estimación de puntos	Intervalo de confianza al 95% Inferior	Superior
Par 1	tasa cardíaca mañana - tasa cardíaca tarde	d de Cohen	2,982	-1,358	-1,962	-,736
		corrección de Hedges	3,107	-1,304	-1,883	-,706

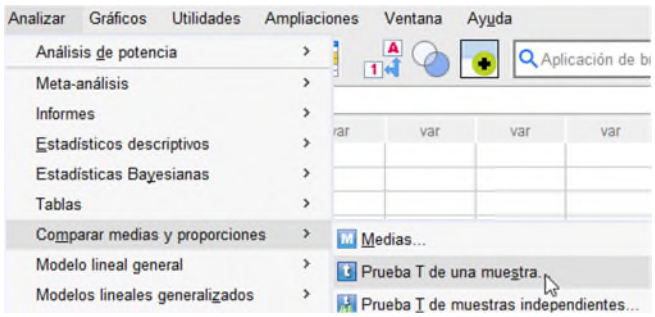
a. El denominador utilizado en la estimación de tamaños del efecto.
La d de Cohen utiliza la desviación estándar de muestra de la diferencia de medias.
La corrección de Hedges utiliza la desviación estándar de muestra de la diferencia de medias, más un factor de corrección.

El valor nos aparece en la columna Estimación de puntos (hay que tomarlo sin signo) que es igual a 1,358, lo que se interpretaría como una medida del tamaño del efecto alta.

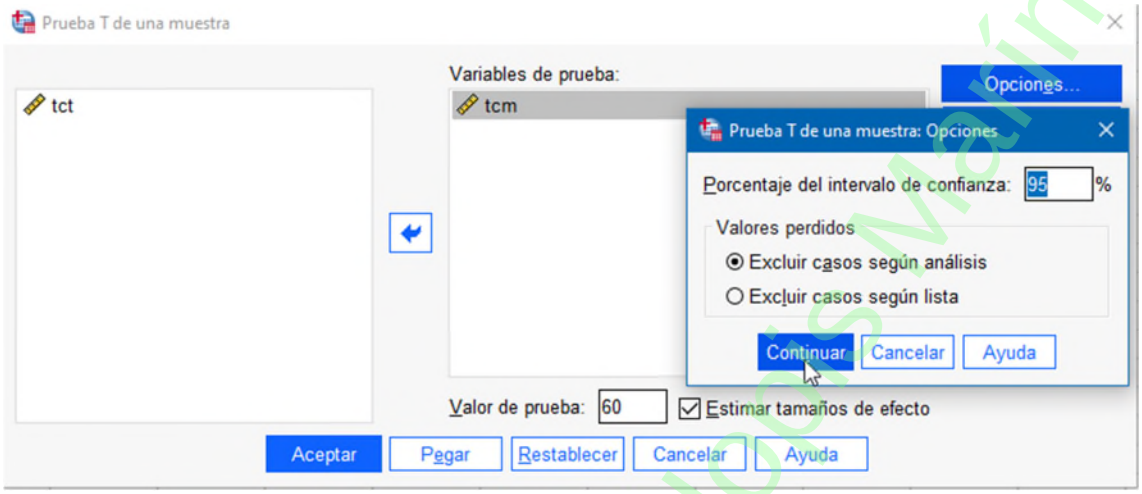
Tamaño del efecto	Intrascendente	Pequeño	Moderado	Alto	Muy alto	Casi perfecto	Perfecto
d	0.0	0.2	0.6	1.2	2.0	4.0	infinite

5.- Test de una sola media

Una última opción es contrastar si una muestra de una variable difiere significativamente de un valor dado o no. Para ello el procedimiento estadístico adecuado que debemos ejecutar es el de *Analizar>Comparar medias y proporciones>Prueba T de una muestra*.



Por ejemplo, si seguimos con los datos anteriores y deseáramos saber si la tasa cardíaca de los 20 sujetos que hemos medido por la mañana difiere por término medio de 60 pulsaciones, este es el procedimiento adecuado que deberíamos ejecutar, seleccionando la variable a contrastar en el cuadro **Contrastar variables** e incluyendo el valor 60 en el cuadro **Valor de la prueba**. El botón **Opciones** permite, al igual que en los procedimientos anteriores, indicar el valor del intervalo de confianza y excluir los casos que se deseen.



Estadísticas para una muestra

	N	Media	Desv. estándar	Media de error estándar
tasa cardíaca mañana	20	64,90	11,706	2,618

Prueba para una muestra

	Valor de prueba = 60						
			Significación			95% de intervalo de confianza de la diferencia	
	t	gl	P de un factor	P de dos factores	Diferencia de medias	Inferior	Superior
tasa cardíaca mañana	1,872	19	,038	,077	4,900	-,58	10,38

Tamaños de efecto de una muestra

		Standardizer[a]	Estimación de puntos	Intervalo de confianza al 95%	
				Inferior	Superior
tasa cardíaca mañana	d de Cohen	11,706	,419	-,044	,871
	corrección de Hedges	12,195	,402	-,042	,836

a. El denominador utilizado en la estimación de tamaños del efecto.
La d de Cohen utiliza la desviación estándar de muestra.
La corrección de Hedges utiliza la desviación estándar de muestra, más un factor de corrección.

En este caso habría que concluir que desde el punto de vista de la significación estadística la media del grupo es igual a 60, en este caso también podemos cuantificar su magnitud (tamaño del efecto), que es igual a 0,419, lo que se interpretaría como una medida del tamaño del efecto pequeña (ver cuadro del contraste anterior). Es decir, aunque hubiésemos sido más liberales en la elección del nivel de significación (p.e. 0,10), la magnitud del efecto encontrada es pequeña.

6.- Algunos supuestos paramétricos

Los contrastes paramétricos son aquellos que cumplen las condiciones siguientes:

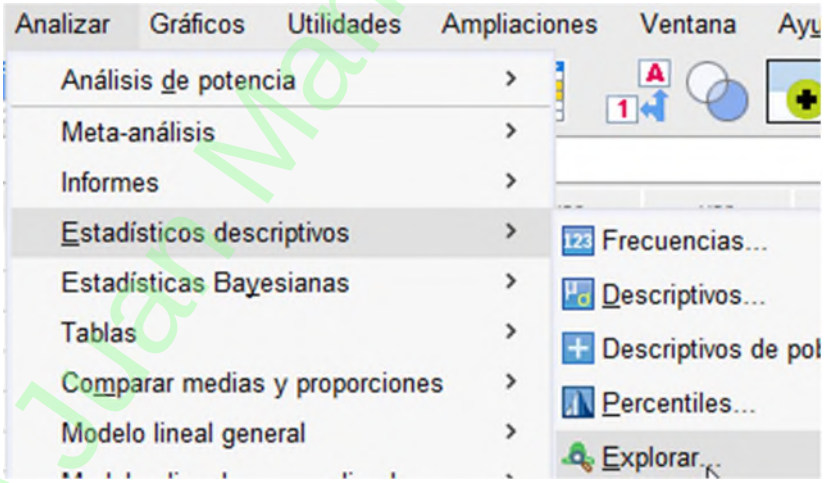
a) Permiten contrastar hipótesis referidas a algún parámetro (t, p, 13, etc.)
b) Requieren el cumplimiento de supuestos sobre las poblaciones de las que se extraen los datos.
c) Analizan datos extraídos con nivel de medida de intervalo o de razón.

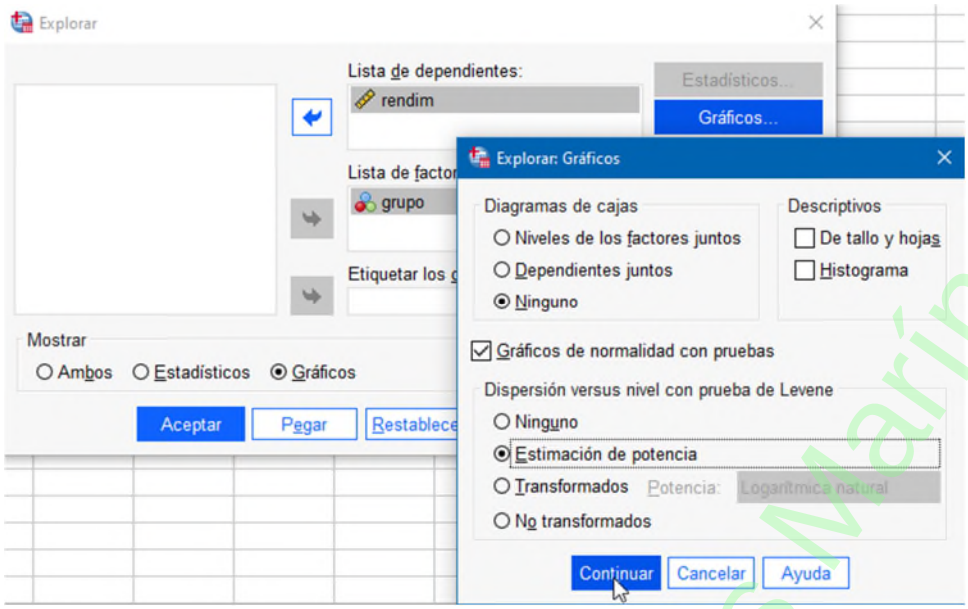
En buena parte de las pruebas de contraste de hipótesis que hemos visto en el presente tema, se necesitan verificar una serie de supuestos que son comunes a las pruebas paramétricas. Alguno de estos supuestos lo hemos comentado a lo largo del tema:

- **Supuesto de homogeneidad de varianzas (homocedasticidad)**. Este supuesto parte de la asunción de que para que los grupos puedan compararse, por ejemplo, a través de sus medias, deben tener la misma variabilidad, y por tanto sus varianzas deben ser iguales. Algunas pruebas ajustan los estadísticos de contraste para el incumplimiento de este supuesto (ver por ejemplo la salida del contraste para dos medias independientes).

- **Supuesto de normalidad**. Se parte de la asunción de que cada condición experimental o cada grupo considerado en el estudio deben proceder de poblaciones con distribución normal. Este supuesto se puede explorar por diferentes métodos gráficos y estadísticos.

Estos dos supuestos pueden ser estudiados a través del procedimiento *Analizar> Estadísticos descriptivos>Explorar*. A continuación, introducimos las variables y marcamos que sólo nos muestre gráficos, pulsando el botón de **Gráficos**, pulsamos las opciones que aparecen en el cuadro siguiente. La salida nos proporciona una serie de estadísticos y gráficos que nos permiten evaluar estos supuestos. **Fichero "2mind.sav"**.





Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	grupo	Estadístico	gl	Sig.	Estadístico	gl	Sig.
rendim	Ruido moderado	,203	9	,200*	,930	9	,477
	Ausencia de ruido	,147	9	,200*	,975	9	,932

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Prueba de homogeneidad de varianza

		Estadístico de Levene	gl1	gl2	Sig.
rendim	Se basa en la media	,007	1	16	,934
	Se basa en la mediana	,000	1	16	1,000
	Se basa en la mediana y con gl ajustado	,000	1	15,529	1,000
	Se basa en la media recortada	,005	1	16	,943

Desde el punto de vista de las pruebas presentadas se podría concluir que los datos cumplen los supuestos de homocedasticidad y normalidad.

Cuando estos supuestos y otros no se cumplen y/o no tenemos un nivel de escala adecuado (inferior a intervalo) existen diferentes alternativas de actuación, entre las que destacamos: (1) No hacer nada, porque ante incumplimientos moderados de los supuestos y bajo ciertas condiciones (tamaño muestral grande, incumplimiento de uno o más supuestos conjuntamente, etc.) algunos tests o pruebas estadísticas se comportan de manera adecuada; (2) otras posible solución es realizar ajustes sobre el test o prueba estadística; (3) una tercera posibilidad es utilizar transformaciones de las variables para conseguir el cumplimiento de los supuestos; y (4) utilizar alternativas no paramétricas, que vamos a ver en el punto siguiente.

7.- Alternativas no paramétricas y análisis categóricos

Los contrastes paramétricos son los que con mayor frecuencia se utilizan en la investigación empírica, pero en diversas ocasiones su aplicación se ve reducida porque, a veces, el cumplimiento de los supuestos no se cumple o porque es necesario trabajar con niveles de medida inferiores a los de intervalo o de razón. Por ello, la estadística ha desarrollado otro conjunto de técnicas que genéricamente se han denominado, por oposición, contrastes no paramétricos. Estos contrastes se caracterizan, por tanto, por:

a) Utilizar datos cuyo nivel de medida es nominal u ordinal (categóricos) o se van a analizar a un nivel de medida nominal u ordinal.
b) Utiliza datos en un nivel de intervalo o de razón, pero la hipótesis planteada no involucra parámetros de las distribuciones poblaciones.
c) El estadístico de contraste utilizado no depende de supuestos específicos de la población en la que se muestrea, salvo de algunos sobre su forma como simetría o continuidad.

Por último, siempre hay que tener en cuenta que, en general, los contrastes no paramétricos son menos potentes que los paramétricos.

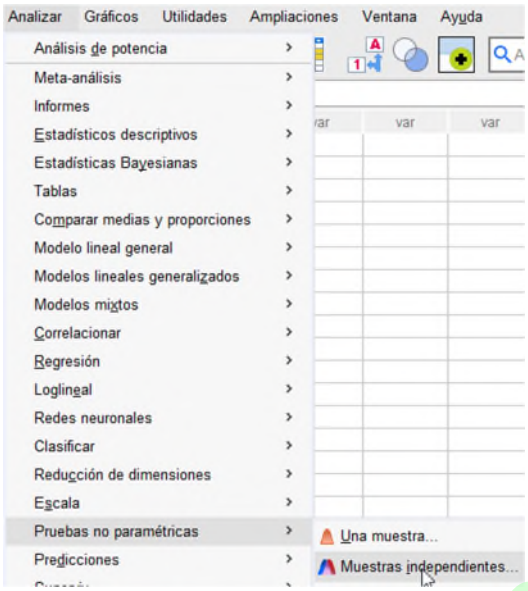
7.1. Análisis de Datos con variables ordinales o cuasi-cuantitativas

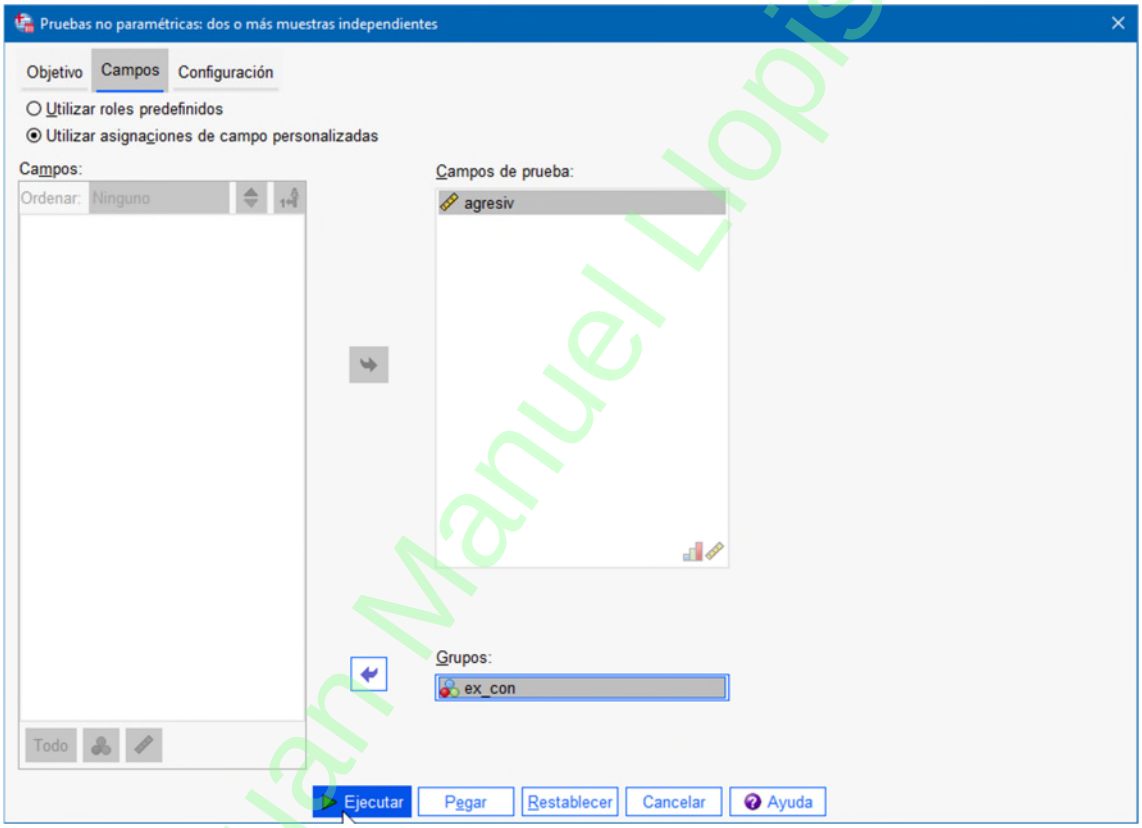
7.1.1. Muestras independientes: U Mann-Whitney

La prueba de Mann-Whitney es una buena alternativa a la prueba t sobre diferencias de medias cuando no se cumplen los supuestos en los que se basa (normalidad y homocedasticidad) o el nivel de medida de los datos es ordinal. Es adecuada en el caso de dos muestras extraídas de dos poblaciones distintas y deseamos averiguar si podemos rechazar la hipótesis de que esas dos poblaciones tienen promedios iguales. Se basa en ordenar las puntuaciones y asignar rangos a esas puntuaciones ordenadas.

Por ejemplo, realizamos un experimento para determinar si un medicamento es efectivo en la reducción de la agresividad. Se forman dos grupos de pacientes, a uno de los cuales se les ha administrado el medicamento (grupo experimental) y al otro no (grupo control). Tras pasar una prueba de agresividad (VD) se han obtenido los resultados siguientes: **Fichero "UMann.sav"**.

¿Podemos afirmar que el tratamiento ha tenido efecto? El procedimiento con SPSS es *Analizar>Pruebas no paramétricas>Muestras independientes*. De las tres pestañas que aparecen, en Campos, colocamos las variables: VD en campos de prueba y VI en grupos y pulsamos ejecutar, el programa detectará nuestros datos y aplicará el análisis adecuado (en este caso U de Mann-Whitney). Si queremos un análisis personalizado, lo elegimos en las pestañas Objetivo y Configuración.





Pruebas no paramétricas

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.[a,b]	Decisión
1	La distribución de agresiv es la misma entre categorías de ex_con.	Prueba U de Mann-Whitney para muestras independientes	,008[c]	Rechace la hipótesis nula.

a. El nivel de significación es de ,050.

b. Se muestra la significancia asintótica.

c. Se muestra la significación exacta para esta prueba.

Prueba U de Mann-Whitney para muestras independientes

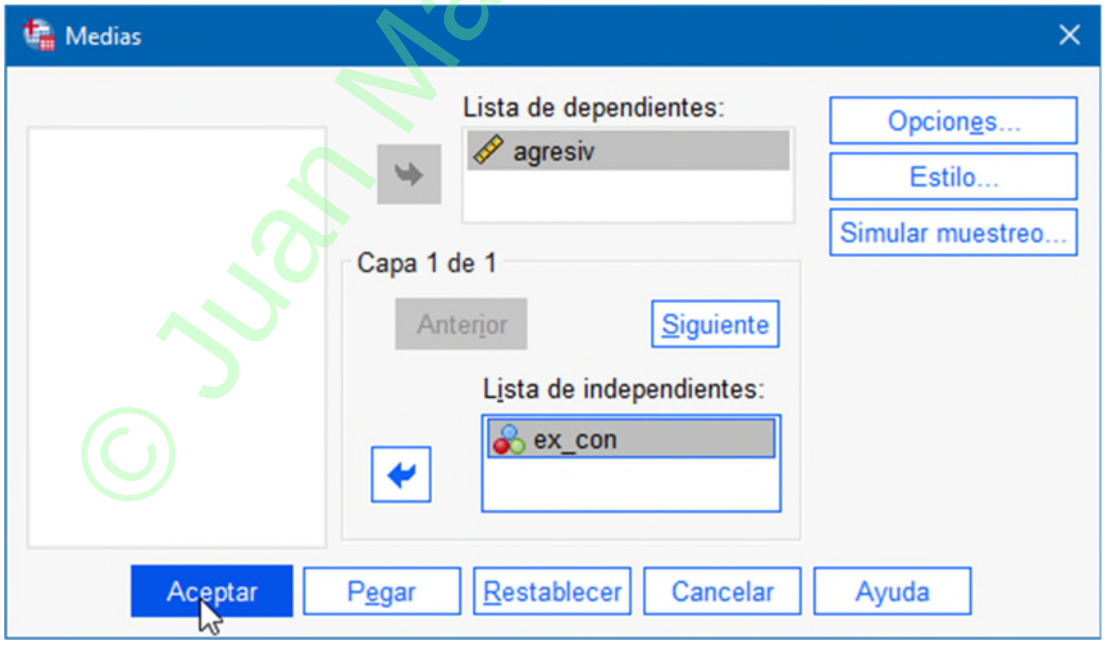
agresiv entre ex_con

Resumen de prueba U de Mann-Whitney de muestras independientes

N total	17
U de Mann-Whitney	62,500
W de Wilcoxon	98,500
Estadístico de prueba	62,500
Error estándar	10,373
Estadístico de prueba estandarizado	2,555
Sig. asintótica (prueba bilateral)	,011
Significación exacta (prueba bilateral)	,008

Se interpreta en la misma línea de las pruebas paramétricas. En este caso concluimos que hay diferencia estadísticamente significativa entre las medias (Estadístico de prueba estandarizado = 2,555; Sig. ≤ 0,05). La decisión de utilizar la significación asintótica o la exacta, depende del tamaño muestral, con muestras pequeñas se debe interpretar la exacta.

Si aparecen diferencias significativas entre las medias de los grupos, para saber cuál es mayor y cual menor, deberemos utilizar el procedimiento *Analizar>Comparar medias y proporciones>Medias.* Colocamos las variables y pulsamos Aceptar.



Informe

agresiv

ex_con	Media	N	Desv. estándar
experimental	9,33	9	3,391
Control	14,62	8	3,462
Total	11,82	17	4,290

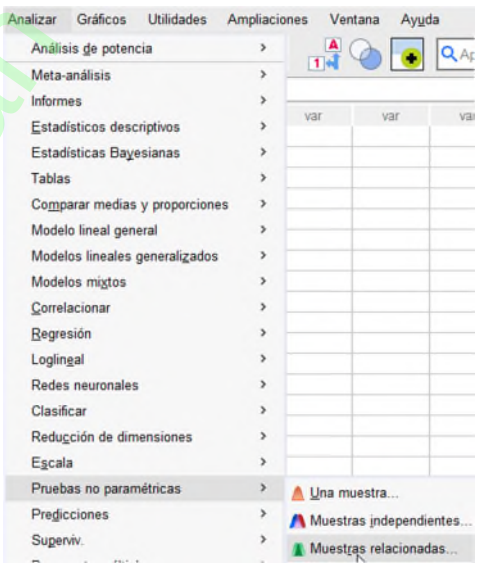
Podemos observar que la media del grupo experimental es significativamente menor que la del grupo control: el medicamento es efectivo en la reducción de la agresividad.

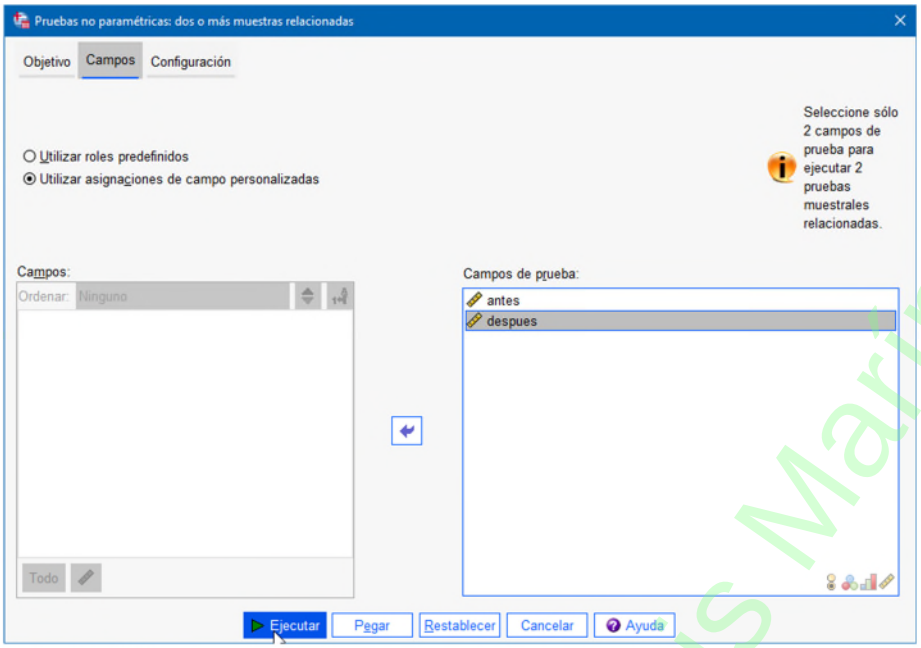
7.1.2. Muestras relacionadas: Wilcoxon y Signos

Permite estudiar si dos muestras relacionadas proceden de dos poblaciones con el mismo promedio. Es una alternativa a la prueba t para muestras relacionadas. Las dos muestras relacionadas se convierten en una sola calculando las diferencias de los pares. Se descartan las diferencias nulas y se calcula la suma del valor absoluto de las diferencias positivas y negativas. Si cualquiera de ellas es muy elevada, posiblemente encontraremos diferencias estadísticamente significativas. Requiere que la VD sea considerada con propiedades cuantitativas, por lo que estrictamente no es una alternativa de análisis categórico, sino de incumplimiento de supuestos. Si la VD es considerada ordinal, la posibilidad que tenemos es realizar la prueba de los signos, que cuenta las diferencias positivas y negativas.

Por ejemplo, un investigador está estudiando si los niveles de ansiedad se reducen después de un curso de entrenamiento en relajación. Para ello selecciona una muestra aleatoria de 10 pacientes y les pide que respondan a un cuestionario de ansiedad antes y después del curso. ¿Permiten estos resultados concluir que la hipótesis del investigador es cierta? **Fichero "Wilcoxon.sav"**.

El procedimiento con SPSS es *Analizar>Pruebas no paramétricas>Muestras relacionadas*. De las tres pestañas que aparecen, en Campos, colocamos las 2 variables en campos de prueba y pulsamos ejecutar, el programa detectará nuestros datos y aplicará el análisis adecuado (en este caso la prueba de Wilcoxon). Si queremos un análisis personalizado, lo elegimos en las pestañas Objetivo y Configuración.





Pruebas no paramétricas

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.[a,b]	Decisión
1	La mediana de diferencias entre antes y despues es igual a 0.	Prueba de rangos con signo de Wilcoxon para muestras relacionadas	,050	Rechace la hipótesis nula.

a. El nivel de significación es de ,050.

b. Se muestra la significancia asintótica.

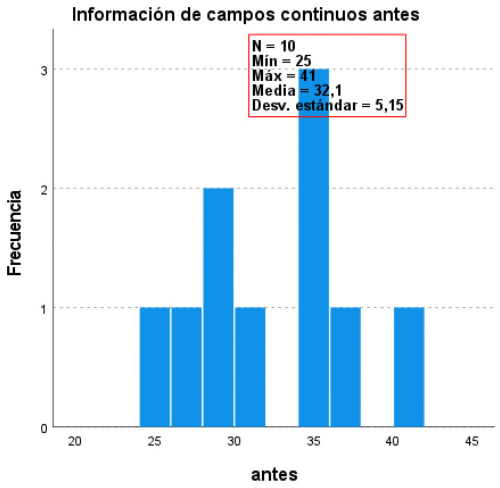
Prueba de rangos con signo de Wilcoxon para muestras relacionadas

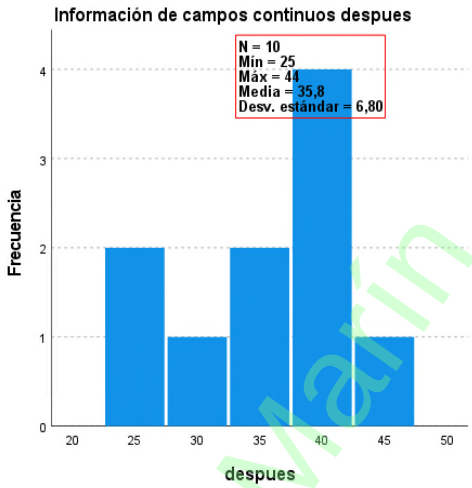
antes, despues

Resumen de prueba de rangos con signo de Wilcoxon para muestras relacionadas

N total	10
Estadístico de prueba	39,000
Error estándar	8,404
Estadístico de prueba estandarizado	1,963
Sig. asintótica (prueba bilateral)	,050

En este caso concluimos que hay diferencias estadísticamente significativas entre las medias (Estadístico de prueba estandarizado = 1,963; Sig. ≤ 0,05). Se concluye que el curso no ha tenido efecto dado que la media de ansiedad después del curso (35,8) es mayor que la media de ansiedad antes del curso (32,1). Las medias aparecen en los gráficos que hay al final de los resultados del análisis:





TEMA - 7

COMPARACIÓN DE MÁS DE DOS MUESTRAS INDEPENDIENTES: ANÁLISIS DE VARIANZA (ANOVA) Y ALTERNATIVAS

1.- Introducción

Los diseños básicos o unifactoriales con más de dos muestras (grupos) independientes son una extensión de los diseños con dos grupos independientes ya visto. Por tanto, esta estructura de investigación queda caracterizada por un único factor o variable independiente que adoptará tres o más niveles. La principal ventaja de este diseño respecto al diseño de dos grupos independientes es que permite obtener una información más precisa acerca de la relación funcional entre la variable independiente y la variable dependiente.

En este tema se va a tratar el Procedimiento de Análisis de la Varianza (ANOVA). El ANOVA es un procedimiento de contraste de hipótesis. Concretamente, es una prueba de contraste de medias que compara simultáneamente dos o más medias. Permite contrastar la hipótesis nula de que las diferencias encontradas entre las medias de diferentes grupos o niveles de una variable independiente no difieren entre sí más de lo que cabría esperar por efecto del azar. Se denomina Análisis de Varianza porque es un procedimiento que permite dividir la variabilidad de la variable dependiente en dos o más componentes, cada uno de los cuales puede ser atribuido a una fuente (variable o factor) identificable. Por tanto, el ANOVA se suele utilizar para decidir si las diferencias que encontramos en nuestros datos, en la variable dependiente, una vez se han aplicado los niveles de la(s) variable(s) independiente(s), pueden ser atribuidas, con el margen de error delimitado por el nivel de significación, al efecto de dicha(s) variable(s) independiente(s), o al efecto de factores aleatorios o azarosos.

2.- Supuestos que deben cumplir los datos para poder analizarlos mediante el modelo de ANOVA

Para poder aplicar correctamente el ANOVA a un conjunto de datos procedentes de la aplicación de un diseño concreto, dichos datos deben satisfacer los siguientes supuestos básicos:

(1) **Normalidad**: las puntuaciones de los diversos grupos en la variable dependiente (VD) se deben distribuir normalmente, lo que implica que son muestras representativas de poblaciones con distribución normal en esa VD. El ANOVA es robusto al incumplimiento de este supuesto, no obstante, si la muestra es pequeña es conveniente evaluarlo.

(2) **Homocedasticidad**: las varianzas poblacionales de los diversos grupos en la VD han de ser homogéneas (iguales), lo que implica que también lo sean las varianzas muestrales. El ANOVA es robusto al incumplimiento de este supuesto con tamaños muestrales iguales en todos los grupos y no muy pequeños. No obstante, es aconsejable evaluarlo. Se puede hacer al mismo tiempo que se ejecuta el Procedimiento para el ANOVA, como se verá en los ejemplos.

(3) **Independencia de las observaciones**: las puntuaciones de los diversos grupos en la VD han de ser independientes, lo que asegura que la razón entre la varianza debida al efecto de la(s) VI(s) y la varianza debida al efecto del error, siga una distribución F de Snedecor con el alfa nominal estipulado y los grados de libertad

asociados al numerador y al denominador de dicha razón. El **ANOVA NO** es robusto al incumplimiento de este supuesto, que se suele incumplir prácticamente siempre que los datos proceden de diseños con medidas repetidas. En estos casos, es necesario tener en cuenta otras opciones que se comentarán en el ejemplo del diseño unifactorial con medidas repetidas.

(4) **Nivel de medida de la VD**: estrictamente, la variable dependiente debe estar medida en una escala de razón o de intervalo.

3.- El Procedimiento Análisis de Varianza en SPSS

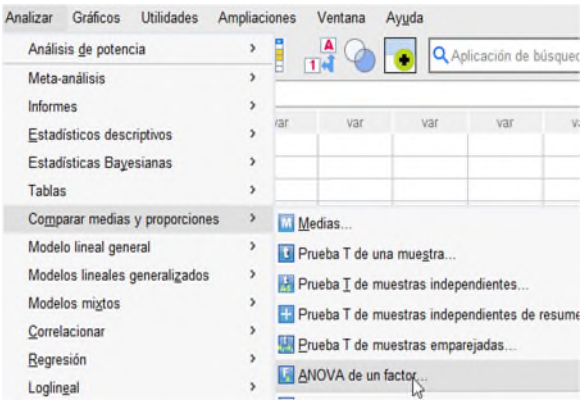
El procedimiento para realizar un ANOVA en SPSS varía en función del tipo de diseño que se haya aplicado en el estudio cuyos datos queramos analizar. Por tanto, en este tema vamos a ver cómo realizar e interpretar un ANOVA para un diseño unifactorial multigrupos (entre grupos).

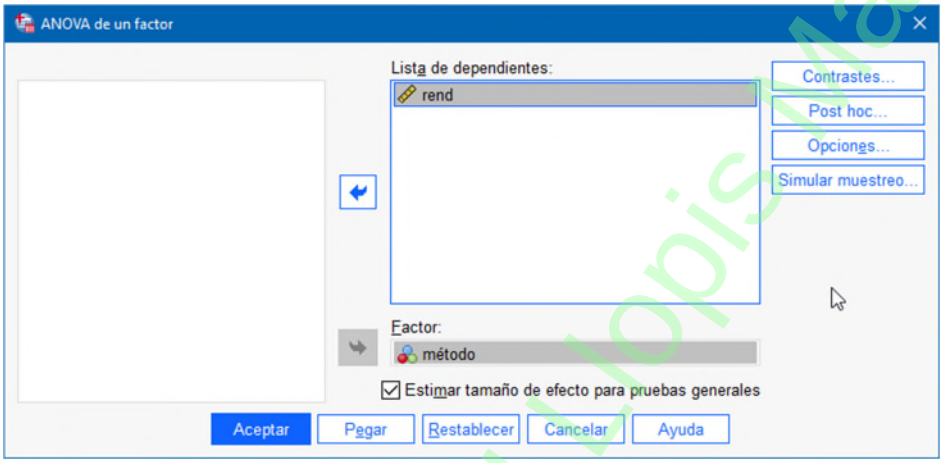
4.- ANOVA para un Diseño Unifactorial Multigrupos (muestras independientes)

4.1. Supuesto práctico. En un Instituto de Secundaria se planteó una discusión en torno a los métodos más adecuados para explicar la asignatura de Geografía. El grupo de profesores más conservador era partidario de continuar con la explicación convencional, basada en el estudio sobre el texto. Los innovadores estaban divididos entre dos opiniones: unos eran partidarios de utilizar el estudio de campo, mientras que otros eran partidarios de la utilización de ordenadores y representaciones gráficas del terreno. Los llamaremos tradicionales, ecológicos e informáticos, respectivamente. Los tradicionales pensaban que su método pedagógico produciría mejores resultados que los restantes. Los innovadores estaban de acuerdo en que sus métodos producirían mejores resultados que el de los tradicionales, pero discrepaban sobre cuál de los dos métodos innovadores sería más efectivo: los ecológicos pensaban que el estudio de campo y los informáticos creían que la utilización del ordenador. Después de ponerse de acuerdo en un método de evaluación del rendimiento, decidieron comparar el efecto de cada método durante un curso completo. Utilizaron para ello una muestra formada por 30 alumnos, que fueron asignados aleatoriamente a cada uno de los tres métodos, formando grupos de 10 alumnos. **Fichero "Multigrupos.sav"**.

4.2. Editor de datos, instrucciones y tipo de análisis

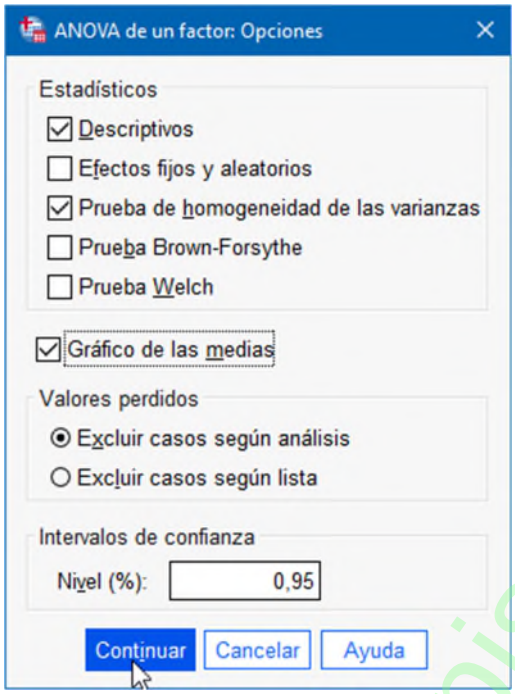
Una vez se han introducido los datos en el Editor de datos, en primer lugar, antes de realizar el ANOVA, es conveniente comprobar si los datos que tenemos cumplen los supuestos necesarios. **Normalidad**: por ejemplo, mediante el Procedimiento *Analizar>Estadísticos Descriptivos>Explorar*. **Homocedasticidad**: se evalúa en el mismo procedimiento que el ANOVA, no es necesario hacerlo antes. **Independencia**: como los datos proceden de un diseño de grupos aleatorios, las puntuaciones se consideran independientes y, por tanto, no es necesario evaluar este supuesto. En caso de querer evaluarlo, se puede utilizar la prueba chi-cuadrado (*Analizar>Estadísticos Descriptivos>Tablas cruzadas*). **Nivel de medida**: la escala de medida que se utiliza es adecuada. Para ejecutar el ANOVA, se elige *Analizar>Comparar medias y proporciones>ANOVA de un factor*.



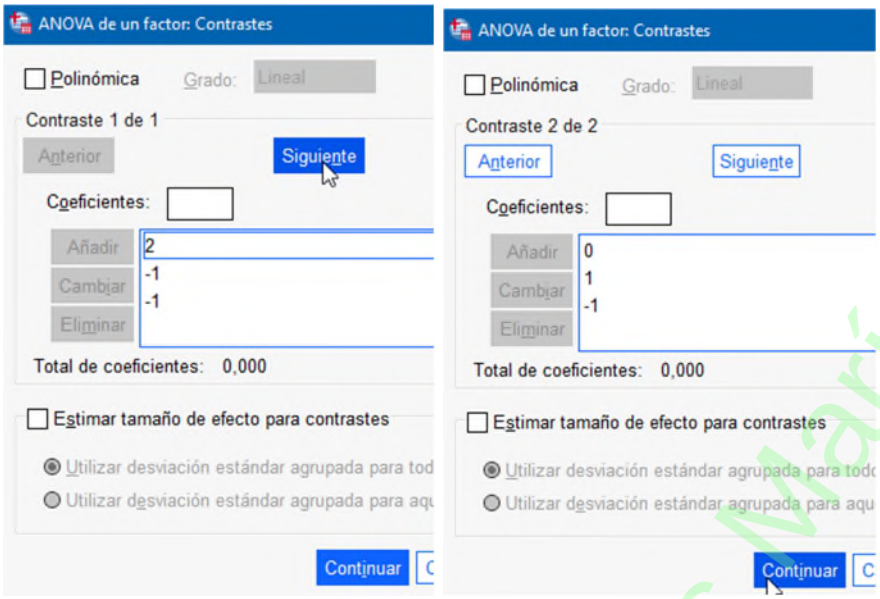


En esta ventana, se deben indicar la variable dependiente y la independiente: introducimos la VD rendimiento (rend) en "Dependientes" y la VI (método) en "Factor". Este procedimiento tiene tres posibilidades además de realizar la prueba global del ANOVA: Contrastes, Post hoc y Opciones.

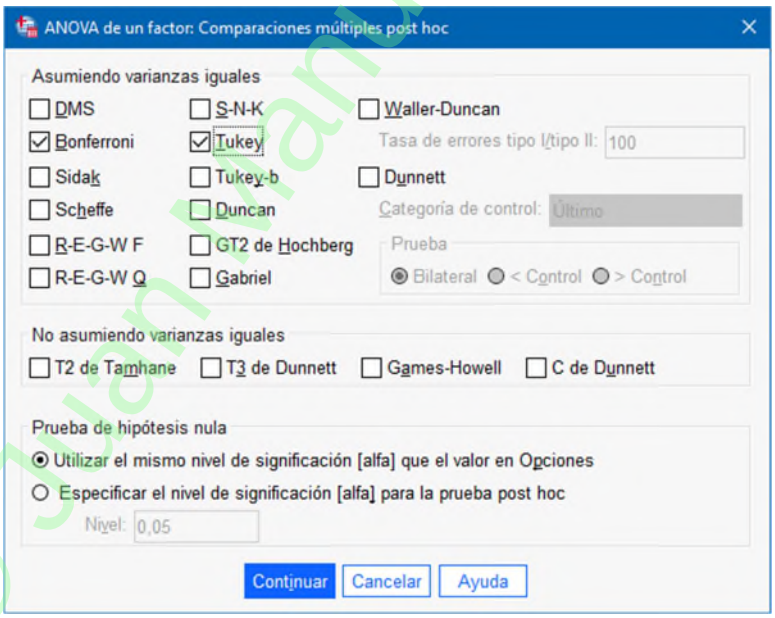
Tanto la opción **Contrastes** como la opción **Post hoc** se utilizan para realizar comparaciones múltiples adicionales a la prueba de ANOVA. La opción **Opciones** permite seleccionar estadísticos descriptivos, la prueba de Levene para comprobar el supuesto de homocedasticidad, gráfico para las medias y permite controlar el tratamiento de valores perdidos, también permite obtener los valores de dos estadísticos (Brown-Forsythe y Welch) que son preferibles al estadístico F de Snedecor cuando no se cumple el supuesto de homocedasticidad.



La opción Contrastes: se utiliza habitualmente si tenemos hipótesis previas acerca del comportamiento de las variables y queremos realizar comparaciones entre los grupos, guiadas por nuestras hipótesis teóricas, o realizar análisis de tendencia si la VI es una variable cuantitativa con intervalos iguales entre sus niveles. Permite realizar varios contrastes, tanto simples como complejos, pudiéndose plantear un conjunto de contrastes ortogonales. Para realizar los contrastes entre medias mediante este procedimiento es necesario asignar coeficientes adecuados a cada contraste que se vaya a realizar. Hay que tener en cuenta que la suma de los coeficientes para cada contraste deberá ser cero. En nuestro ejemplo, podemos plantear dos contrastes ortogonales, por ejemplo, el primero, el método tradicional frente al método ecológico e informático considerados conjuntamente (coeficientes: 2, -1, -1), y el segundo, el método ecológico frente al informático (coeficientes: 0, 1, -1). Los coeficientes se deben introducir uno a uno respetando el orden establecido para el contraste y pulsar Añadir después de cada entrada. El orden de entrada es importante porque se corresponde con el orden de las categorías de la variable independiente o factor. Una vez se han introducido los coeficientes para el primer contraste, pulsar Siguiente para introducir los correspondientes al segundo contraste. Después Continuar para volver a la ventana principal.



La opción Post hoc: se utiliza habitualmente cuando no partimos de hipótesis claras acerca del comportamiento de las variables y, una vez que el ANOVA ha indicado que existen diferencias estadísticamente significativas entre los grupos, queremos realizar todas las comparaciones posibles entre ellos para obtener la máxima información. La ventana de comparaciones múltiples permite elegir entre distintos estadísticos para realizar las comparaciones. Casi todos realizan comparaciones simples, entre pares de medias, a excepción de la prueba de Scheffé. Es aconsejable seleccionar dos de ellos y comprobar si coinciden los resultados, en este sentido, dos buenas opciones pueden ser la prueba de Bonferroni y la de Tukey, para el caso de que no se incumpla el supuesto de homocedasticidad, y las pruebas T3 o C de Dunnett, si se incumple.



4.3. Resultados e Interpretación

En primer lugar, aparecen los descriptivos y la prueba para evaluar el supuesto de igualdad de varianzas que solicitamos en la ventana Opciones.

Descriptivos

Rendimiento

	N	Media	Desviación estándar	Error estándar	95% de intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
Tradicional	10	9,80	3,910	1,236	7,00	12,60	4	15
Ecológico	10	29,80	3,736	1,181	27,13	32,47	25	35
Informático	10	33,50	4,767	1,507	30,09	36,91	25	40
Total	30	24,37	11,324	2,068	20,14	28,60	4	40

Pruebas de homogeneidad de varianzas

		Estadístico de Levene	gl1	gl2	Sig.
Rendimiento	Se basa en la media	,119	2	27	,888
	Se basa en la mediana	,136	2	27	,874
	Se basa en la mediana y con gl ajustado	,136	2	24,022	,874
	Se basa en la media recortada	,119	2	27	,888

Como se puede ver en la tabla, el resultado de la prueba indica que se puede asumir la igualdad de varianzas de las puntuaciones de los tres grupos implicados.

A continuación, aparece la tabla resumen del ANOVA para las variables que hemos indicado. El resultado indica que existen diferencias estadísticamente significativas entre los grupos.

ANOVA

Rendimiento

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	3251,267	2	1625,633	93,847	<,001
Dentro de grupos	467,700	27	17,322		
Total	3718,967	29			

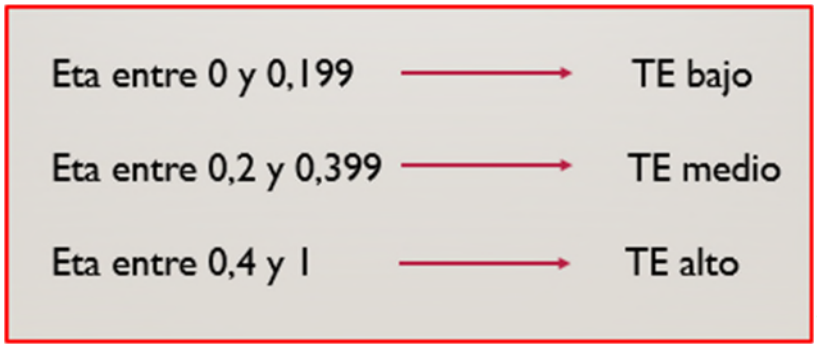
A continuación, tenemos el cálculo del tamaño del efecto:

Tamaños de efecto ANOVA[a]

		Estimación de puntos	Intervalo de confianza al 95%	
			Inferior	Superior
Rendimiento	Eta cuadrado	,874	,750	,913
	Epsilon cuadrado	,865	,732	,907
	Omega cuadrado efecto fijo	,861	,725	,904
	Omega cuadrado efecto aleatorio	,756	,569	,825

a. Eta cuadrado y Epsilon cuadrado se estiman basándose en el modelo de efecto fijo.

Una guía que puede orientar a partir de qué valores se puede considerar un tamaño del efecto bajo, medio o alto, es la siguiente (Cohen, 1988):



En este ejemplo, tenemos un tamaño del efecto alto.

Seguidamente, la salida proporciona la información correspondiente a la opción Contrastes: una tabla con los coeficientes que hemos especificado para los contrates ortogonales que deseábamos realizar, y otra con el resultado de dichos contrastes. Recordemos que los contrastes que especificamos mediante los coeficientes eran: el primero, el método tradicional frente al método ecológico e informático considerados conjuntamente (coeficientes: 2, -1, -1), y el segundo, el método ecológico frente al informático (coeficientes: 0, 1, -1).

Coeficientes de contraste

	Método de enseñanza		
Contraste	Tradicional	Ecológico	Informático
1	2	-1	-1
2	0	1	-1

Pruebas de contraste

		Contraste	Valor de contraste	Error estándar	t	gl	Sig. (bilateral)
Rendimiento	Asume varianzas iguales	1	-43,70	3,224	-13,555	27	<,001
		2	-3,70	1,861	-1,988	27	,057
	No se asume varianzas iguales	1	-43,70	3,128	-13,971	19,353	<,001
		2	-3,70	1,915	-1,932	17,027	,070

Dado que se ha asumido igualdad de varianzas, los resultados adecuados son los de la primera fila de la tabla. Estos resultados indican que podemos rechazar la H_0 para el primer contraste, pero no para el segundo. Es decir, que existen diferencias estadísticamente significativas entre la puntuación media de los alumnos a los que se les ha aplicado el método tradicional y la media conjunta de los alumnos a los que se les han aplicado los otros dos métodos. En concreto, si miramos las medias, podemos ver que el rendimiento medio del grupo tradicional es menor que la media conjunta de los otros dos grupos. Sin embargo, el resultado del segundo contraste indica que las diferencias entre las medias de los grupos ecológico e informático no resultan estadísticamente significativas, por lo que esa diferencia debe ser atribuida al efecto del azar. No podemos, por tanto, decir que uno de estos dos métodos produzca un mayor rendimiento que el otro, aunque sí que, en conjunto, producen un mejor rendimiento que el método tradicional.

Los resultados que aparecen a continuación son las comparaciones Post hoc, que proporcionan información muy similar a los contrastes anteriores. De hecho, llegamos a las mismas conclusiones con unos u otros, aunque por diferentes métodos.

Pruebas post hoc

Comparaciones múltiples

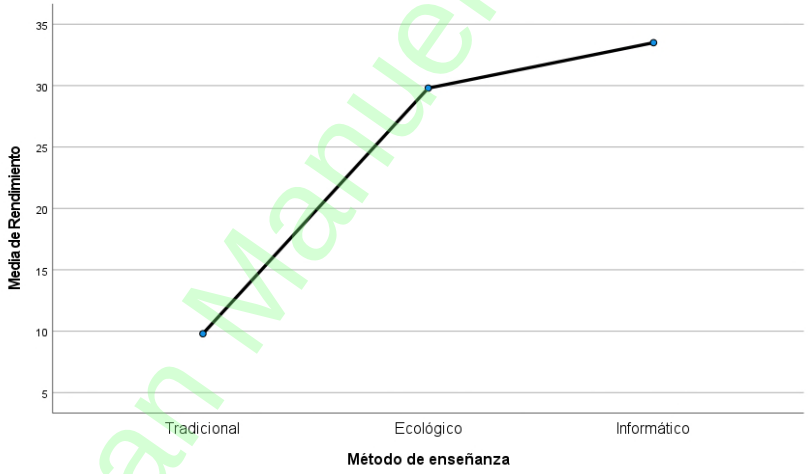
Variable dependiente: Rendimiento

	(I) Método de enseñanza	(J) Método de enseñanza	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95% Límite inferior	Límite superior
HSD Tukey	Tradicional	Ecológico	-20,000*	1,861	<0,001	-24,61	-15,39
		Informático	-23,700*	1,861	<0,001	-28,31	-19,09
	Ecológico	Tradicional	20,000*	1,861	<0,001	15,39	24,61
		Informático	-3,700	1,861	0,135	-8,31	0,91
	Informático	Tradicional	23,700*	1,861	<0,001	19,09	28,31
		Ecológico	3,700	1,861	0,135	-0,91	8,31
Bonferroni	Tradicional	Ecológico	-20,000*	1,861	<0,001	-24,75	-15,25
		Informático	-23,700*	1,861	<0,001	-28,45	-18,95
	Ecológico	Tradicional	20,000*	1,861	<0,001	15,25	24,75
		Informático	-3,700	1,861	0,171	-8,45	1,05
	Informático	Tradicional	23,700*	1,861	<0,001	18,95	28,45
		Ecológico	3,700	1,861	0,171	-1,05	8,45

*. La diferencia de medias es significativa en el nivel 0.05.

Como se puede ver en la tabla, ambas pruebas arrojan los mismos resultados. Concretamente, vemos que las diferencias estadísticamente significativas se encuentran entre el grupo control (método tradicional) y los dos grupos experimentales (métodos ecológico e informático). Sin embargo, no existen diferencias estadísticamente significativas entre ambos grupos experimentales. Este resultado está indicando que ambos métodos de enseñanza no difieren significativamente entre sí en cuanto a su efectividad para aumentar el rendimiento.

Finalmente, tenemos el gráfico de medias que también solicitamos en Opciones.



5.- Alternativas para datos categóricos

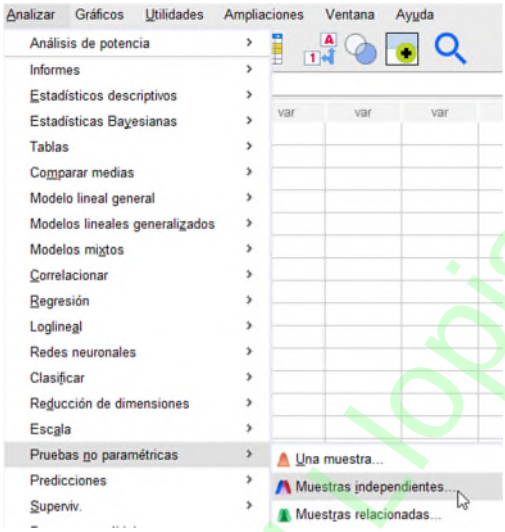
5.1. Contrastes para más de dos muestras independientes (datos ordinales mínimo)

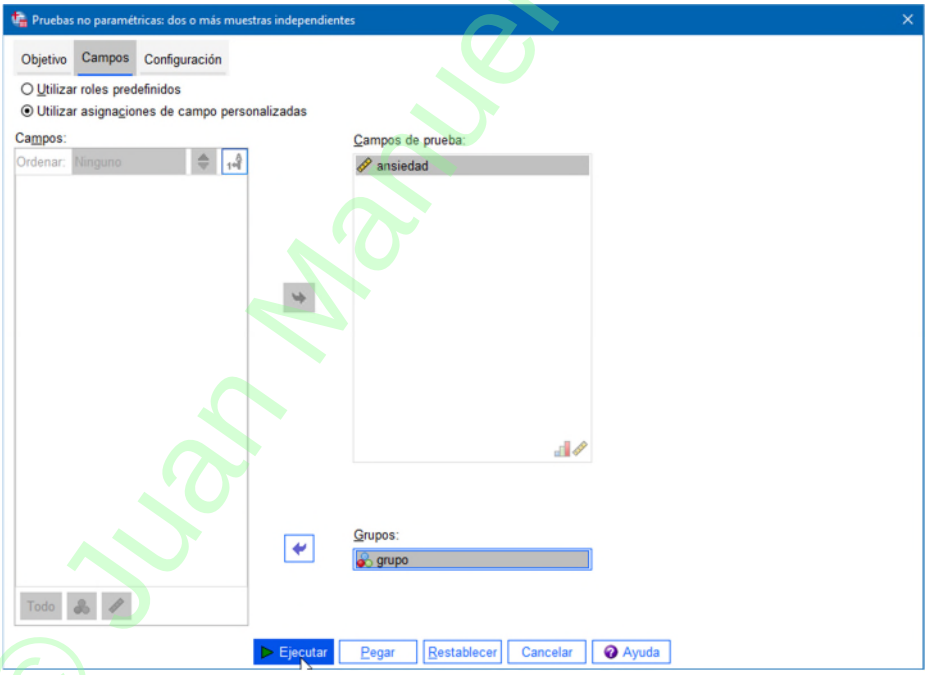
Es similar al ANOVA de un factor completamente aleatorizado. Las ventajas fundamentales respecto al ANOVA son que no necesita establecer supuestos sobre las poblaciones (normalidad y homocedasticidad) y que permite trabajar con datos ordinales.

Ejemplo: Un psicólogo estaba interesado en estudiar el nivel de ansiedad que sufren las personas que han intentado suicidarse frente a otros grupos de personas. Para ello eligió a 9 pacientes que habían intentado suicidarse (S), a 11 pacientes neurótico

obsesivos (N-O) y a 10 sujetos considerados normales (N). A todos ellos les pasó la misma escala de ansiedad. ¿Podemos considerar la hipótesis de que existen diferencias en ansiedad entre los tres grupos? **Fichero: "krus-W.sav"**.

Para responder a esta pregunta con ayuda del SPSS elegimos: *Analizar>Pruebas no paramétricas>Muestras independientes.* De las tres pestañas que aparecen, en Campos, colocamos las variables: VD en campos de prueba y VI en grupos y pulsamos ejecutar, el programa detectará nuestros datos y aplicará el análisis adecuado (en este caso ANOVA de 1 factor de Kruskal-Wallis). Si queremos un análisis personalizado, lo elegimos en las pestañas Objetivo y Configuración.





La salida es:

Pruebas no paramétricas

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.[a,b]	Decisión
1	La distribución de ansiedad es la misma entre categorías de grupo.	Prueba de Kruskal-Wallis para muestras independientes	<,001	Rechace la hipótesis nula.

a. El nivel de significación es de ,050.

b. Se muestra la significancia asintótica.

Prueba de Kruskal-Wallis para muestras independientes

ansiedad entre grupo

Resumen de prueba Kruskal-Wallis de muestras independientes

N total	30
Estadístico de prueba	22,269[a]
Grado de libertad	2
Sig. asintótica (prueba bilateral)	<,001

a. Las estadísticas de prueba se ajustan para empates.

Lo que nos lleva a concluir que existen diferencias entre los grupos considerados. La novedad sobre los procedimientos no paramétricos vistos anteriormente se establece en que debemos explorar entre qué grupos se encuentran esas diferencias. El resultado de la comparación aparece en la siguiente tabla:

Comparaciones por parejas de grupo

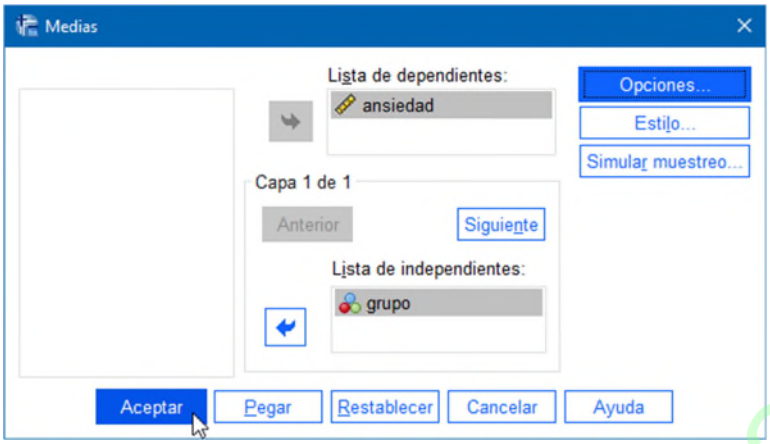
Sample 1-Sample 2	Estadístico de prueba	Desv. Error	Desv. Estadístico de prueba	Sig.	Sig. ajustada[a]
normales-suicidas	8,372	4,044	2,071	,038	,115
normales-neurótico-obsesivos	18,105	3,845	4,708	<,001	,000
suicidas-neurótico-obsesivos	-9,732	3,956	-2,460	,014	,042

Cada fila prueba la hipótesis nula que las distribuciones de la Muestra 1 y la Muestra 2 son iguales.
Se visualizan las significaciones asintóticas (pruebas bilaterales). El nivel de significación es de ,050.

a. Los valores de significación se han ajustado mediante la corrección Bonferroni para varias pruebas.

Podemos observar que existen diferencias significativas entre normales y neurótico-obsesivos y entre suicidas y neurótico-obsesivos (Sig. Ajustada < 0,05). La prueba de Kruskal-Wallis aplica automáticamente la corrección Bonferroni a las comparaciones post hoc.

Si aparecen diferencias significativas entre las medias de los grupos, para saber cuál es mayor y cual menor, deberemos utilizar el procedimiento *Analizar>Comparar medias y proporciones>Medias* colocar las variables y pulsar Aceptar.



Informe

ansiedad

grupo	Media	N	Desv. Desviación
suicidas	28,22	9	5,472
neurótico-obsesivos	43,27	11	8,742
normales	15,30	10	6,881
Total	29,43	30	13,831

Podemos observar que en la diferencia entre normales y neuróticos-obsesivos, la media de ansiedad es significativamente mayor en estos últimos.

En la diferencia entre suicidas y neuróticos-obsesivos, la media de ansiedad también es significativamente mayor en los neurótico-obsesivos.

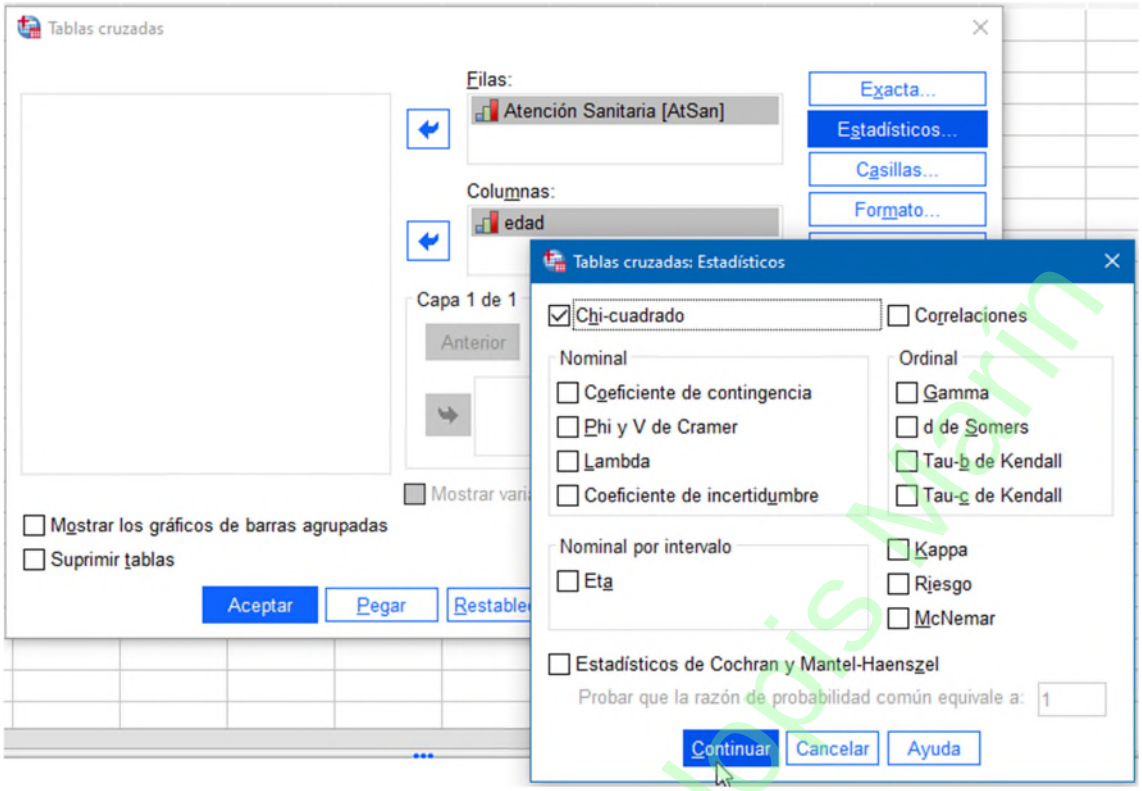
En conclusión, los sujetos neurótico-obsesivos tienen una media de ansiedad significativamente mayor que los suicidas y los normales. Entre suicidas y normales no hay diferencias significativas en su nivel de ansiedad.

5.2. Contrastes para más de dos muestras independientes (Datos nominales mínimo)

Nos permite contrastar hipótesis sobre la igualdad de más de dos proporciones para muestras independientes.

Ejemplo: Para estudiar si la actitud hacia la atención sanitaria va cambiando con la edad se tomaron tres muestras aleatorias de 40 sujetos de distintas edades categorizadas (alta, media y baja). Todos los sujetos respondieron a la pregunta ¿Está conforme con la atención sanitaria que recibe? En una escala de 1 a 4 (nada, poco, bastante, mucho). Tras los resultados, ¿se puede afirmar que las distintas edades difieren en su actitud hacia la atención sanitaria? **Fichero "chi_c3.sav".**

Para responder a esta pregunta con ayuda del SPSS; *Analizar>Estadísticos Descriptivos>Tablas cruzadas*. Introducimos una de las variables en filas y la otra en columnas, y pulsamos en el botón de **Estadísticos**, **Chi-cuadrado**.



La salida básica es:

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	29,545[a]	6	<0,001
Razón de verosimilitud	31,886	6	<0,001
Asociación lineal por lineal	3,092	1	0,079
N de casos válidos	120		

a. 3 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 4,00.

Lo que nos lleva a concluir que existen diferencias entre los grupos considerados. Como hemos indicado anteriormente, debemos explorar entre qué grupos se encuentran esas diferencias.

En este caso no tenemos pruebas no paramétricas a posteriori, por lo que procederemos a realizar tantas pruebas Chi-Cuadrado como comparaciones se deseen realizar. Teniendo en cuenta que hay que corregir el nivel de significación en función de, por ejemplo, la corrección de Bonferroni (ver recuadro posterior) que vuelve a establecerse en 0,05/3 =0,0167. Si ejecutamos *Analizar>Estadísticos Descriptivos>Tablas cruzadas*, para cada una de las comparaciones, obtenemos que existen diferencias entre el grupo bajo y medio (Edad = 1 y 2), pero no entre los grupos bajo y alto (Edad = 1 y 3), ni entre los grupos medio y alto (Edad = 2 y 3).

Error de tipo I y tasa de error de tipo I en comparaciones múltiples:

Cuando sometemos a prueba una hipótesis nula, como en el ANOVA, la probabilidad de cometer un error de Tipo I es igual al valor que estipulamos para α, es decir, 0,05 (lo que se conoce como alfa nominal: α_N) Sin embargo, cuando contrastamos varias hipótesis nulas (por ejemplo, en los contrastes post hoc para datos nominales), la probabilidad de cometer al menos un error de Tipo I en una de ellas aumenta, pasando a ser mayor de 0,05.

Podemos controlar la Tasa de Error Tipo I de un experimento con la denominada Corrección de Bonferroni. Es un procedimiento muy sencillo que consiste en aplicar en cada comparación individual un nivel de alfa, que es el cociente entre el α_N final que se quiere asumir y el número de comparaciones que realiza:

$$\alpha_{PH} = \alpha_N / C$$

De esta forma, en el caso de formular 3 contrastes post hoc para que el α_N final se mantenga en 0,05, en cada comparación individual se tendrá que asumir el siguiente valor de alfa:

$$\alpha_{PH} = \alpha_N / C = 0,05 / 3 = 0,0167$$

Cuando la variable tiene 3 grupos, el número de comparaciones es 3 (C = 3). Cuando la variable tiene 4 grupos, el número de comparaciones es 6 (C = 6). Cuando la variable tiene 5 grupos, el número de comparaciones es 10 (C = 10). $C = \frac{G \cdot (G-1)}{2}$ donde G es el número de grupos y C el número de comparaciones.

Para poder realizar las comparaciones, debemos seleccionar los niveles de la variable edad. *Datos>Seleccionar casos*. Luego debemos establecer la opción "Si se satisface la condición" siguientes:

edad = 1 OR edad = 2
edad = 1 OR edad = 3
edad = 2 OR edad = 3

Después de realizar cada una de las selecciones, ejecutamos *Analizar>Estadísticos Descriptivos>Tablas cruzadas*, eligiendo el estadístico Chi-Cuadrado. Los resultados son:

Pruebas de chi-cuadrado - Edad = 1 y 2

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	27,849[a]	3	<0,001
Razón de verosimilitud	31,018	3	<0,001
Asociación lineal por lineal	13,298	1	<0,001
N de casos válidos	80		

a. 4 casillas (50,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,50.

Pruebas de chi-cuadrado - Edad = 1 y 3

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	9,590[a]	3	0,022
Razón de verosimilitud	10,150	3	0,017
Asociación lineal por lineal	3,463	1	0,063
N de casos válidos	80		

a. 2 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,50.

Pruebas de chi-cuadrado - Edad = 2 y 3

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	7,096[a]	3	0,069
Razón de verosimilitud	7,232	3	0,065
Asociación lineal por lineal	2,782	1	0,095
N de casos válidos	80		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 5,00.

TEMA - 8

COMPARACIÓN DE MÁS DE DOS MUESTRAS DEPENDIENTES: ANÁLISIS DE VARIANZA CON MEDIDAS REPETIDAS (ANOVA MR) Y ALTERNATIVAS

1.- Características fundamentales de los diseños con medidas repetidas

En los diseños de investigación con medidas repetidas, denominados también "intrasujetos", a diferencia de los diseños de grupos que se han visto en el tema anterior, cada uno de los sujetos o participantes en la investigación es sometido a todos los niveles de las variables independientes. Por tanto, cada sujeto tendrá tantas puntuaciones como niveles tenga la variable independiente del estudio, en el caso de un diseño básico o unifactorial, o como condiciones experimentales se hayan generado, en el caso de un diseño factorial. Es decir, tendremos más puntuaciones que sujetos participantes en el estudio en cuestión.

2.- Análisis de Varianza para un Diseño Unifactorial con Medidas Repetidas

Al igual que en los diseños anteriores, antes de aplicar el modelo de análisis de la varianza a los datos de un determinado estudio, es necesario evaluar los supuestos de aplicación. Los supuestos de Normalidad y Homogeneidad de varianzas (homocedasticidad) se tienen que evaluar, para este tipo de diseños, previamente a la ejecución del procedimiento que realiza el ANOVA, como se indicará posteriormente en el ejemplo. Nos pararemos en el supuesto de independencia de las observaciones, ya que es muy importante en los diseños MR.

En el caso de los diseños con medidas repetidas, dado que son los mismos sujetos los que reciben cada una de las condiciones experimentales, es muy probable que aparezca correlación entre sus puntuaciones y, por tanto, que se incumpla este supuesto (independencia de las observaciones), que es el más importante del ANOVA. El incumplimiento de este supuesto repercute gravemente en los resultados arrojados por la prueba F, ya que no es robusta ante observaciones correlacionadas. Concretamente, la correlación entre las puntuaciones provoca un sesgo positivo en la prueba F, aumentando la probabilidad de cometer errores tipo I. Es decir, aumenta la probabilidad de encontrar resultados estadísticamente significativos cuando, realmente, no lo son. Sin embargo, se ha demostrado que para que el comportamiento de la prueba F sea adecuado, es suficiente con el cumplimiento del denominado supuesto de esfericidad. Así que, en la práctica, el supuesto de independencia de los errores es reemplazado por la asunción o supuesto de esfericidad[1].

Por lo tanto, para analizar los datos procedentes de un diseño MR mediante el modelo univariado de Análisis de Varianza (ANOVA), será necesario comprobar si en dichos datos se cumple el supuesto de esfericidad, siempre que la(s) VI(s) implicada(s) tengan más de dos niveles[2].

[1] Para que se cumpla este supuesto, las varianzas de las puntuaciones de diferencia entre pares de condiciones experimentales o niveles deben ser homogéneas (es decir, que las diferencias existentes entre ellas tienen que ser debidas a factores aleatorios, y no a la existencia de un efecto sistemático en las respuestas). La homogeneidad de las varianzas de las diferencias implica que la matriz de varianzas-covarianzas presenta una forma determinada denominada esfericidad o circularidad, de ahí el nombre del supuesto.

[2] Ya que, si sólo se han realizado dos medidas por sujeto, habrá únicamente una puntuación de diferencia y la presencia o no de esfericidad no cabe planteársela.

Uno de los test más utilizados para evaluar el supuesto de esfericidad es la W de Mauchly (1940). Esta prueba la aporta por defecto la salida que el SPSS proporciona al analizar datos con medidas repetidas mediante el procedimiento *Analizar>Modelo Lineal General>Medidas Repetidas*.

Interpretación del test de Mauchly: para que se pueda asumir el supuesto de esfericidad, se debe mantener la H_0 (esfericidad), entonces, los datos se pueden analizar mediante el modelo univariado de ANOVA sin correr el riesgo de obtener un valor sesgado de la prueba F. Sin embargo, si se rechaza la H_0, el supuesto de esfericidad no se puede mantener y, por tanto, el modelo univariado de ANOVA no es apropiado para analizar los datos[3].

En el caso de que NO se pueda mantener el supuesto de esfericidad, se han propuesto varias opciones para analizar los datos, tres de las cuales las proporciona el procedimiento *Analizar>Modelo Lineal General>Medidas Repetidas*, por defecto:

(1) **Utilizar comparaciones específicas (análisis de tendencias)** entre los niveles de la(s) variables independientes en lugar de la prueba general del ANOVA; ya que estas comparaciones no requieren la esfericidad. Esta opción puede ser adecuada sólo cuando la VI sea cuantitativa, los niveles están igual de espaciados entre sí e interese conocer el patrón de la evolución de los valores de la VD. Se suele utilizar casi exclusivamente cuando la VI es el paso del tiempo. El programa SPSS las denomina Pruebas de contrastes intrasujetos.

(2) **Mantener el modelo univariado de ANOVA** con ajustes sobre los grados de libertad asociados al numerador (variabilidad debida al efecto del tratamiento) y al denominador (variabilidad debida al error) de la prueba F, compensando de esta manera el sesgo positivo de la prueba F bajo ausencia de esfericidad. Los ajustes se realizan mediante el parámetro épsilon. El programa SPSS proporciona tres valores para épsilon, que aparecen en la salida en la misma tabla que el test de Mauchly. También proporciona las tablas resumen de los ANOVAS ajustados junto con los resultados para cuando se asume la esfericidad, en la tabla denominada Pruebas de los efectos intrasujetos. Se recomienda el de Greenhouse-Geisser, ya que habitualmente proporciona valores menos extremos.

(3) **Utilizar una aproximación multivariada sobre las medidas repetidas (MANOVA)**, que no requiere el cumplimiento del supuesto de esfericidad. Bajo este modelo, cada una de las medidas que se registran en cada sujeto de forma repetida, se consideran como una variable dependiente diferente. Esta opción aparece la primera en la salida, y se denomina Contrastes multivariados.

De las tres opciones, las más utilizadas son la segunda y la tercera. Sin embargo, una aproximación no es mejor que otra como regla general.

[3] Contrasta la hipótesis nula de que las diferencias existentes entre las varianzas de las puntuaciones de diferencia son debidas a factores aleatorios (efectos del muestreo) y, por tanto, dichas varianzas se pueden considerar homogéneas. Si se acepta la H_0 ($p > 0,05$), entonces los datos se pueden analizar mediante el modelo univariado de ANOVA sin correr el riesgo de obtener un valor sesgado de la prueba F.

Criterios para decidir entre la aproximación univariada o la aproximación multivariada

El modelo multivariado tiene como ventaja respecto al univariado los ajustes sobre los grados de libertad que asegura, matemáticamente, la igualdad entre el error de tipo I y el α nominal fijado a priori, mientras que el univariado sólo aproxima dichos valores. Sin embargo, el modelo multivariado requiere que se cumpla el supuesto de normalidad multivariada para que los resultados sean correctos.

Por otra parte, hay que tener en cuenta el tamaño muestral. Así, Maxwell y Delaney (1990) recomiendan la aproximación multivariada siempre que se cumpla el supuesto de normalidad multivariada y el número de sujetos no sea excesivamente pequeño, en concreto, n debería superar en 10 el número de medidas repetidas para que este modelo tenga una elevada potencia (otros autores recomiendan superarlo en al menos 20). Pero si n es escasamente superior al número de medidas repetidas, entonces el modelo univariado tiene mayor potencia que el multivariado.
Finalmente, cuando se cumple el supuesto de esfericidad, la aproximación univariada tiene mayor potencia que la multivariada y, por tanto, es más aconsejable.

2.1.- Supuesto práctico

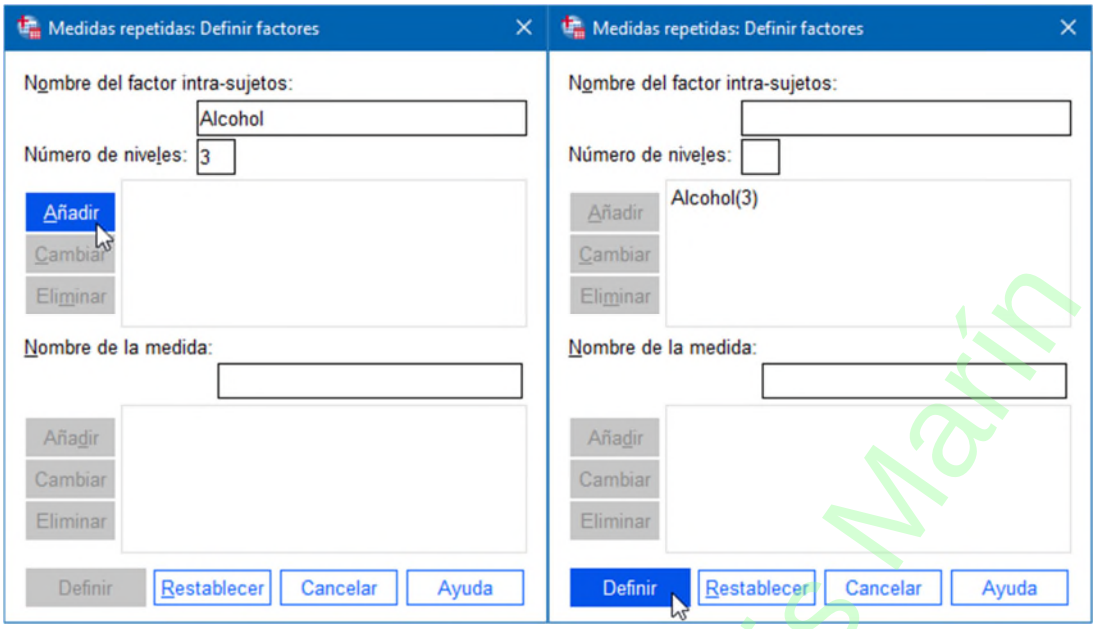
En un centro psicotécnico, se quiere llevar a cabo un estudio para analizar cómo diferentes niveles de alcoholemia en sangre pueden afectar a la tarea de conducción de un automóvil a través de un simulador nuevo que han adquirido. Concretamente, están interesados en saber si aumentos de 0,5 gramos de alcohol en sangre pueden disminuir el número de aciertos en la tarea a realizar. Para llevar a cabo el estudio, seleccionaron una muestra de 24 sujetos, todos ellos con 5 años de carnet. El nivel de alcoholemia se manipuló a tres niveles: 0 gramos, 0,5 gramos y 1 gramo. Cada uno de los sujetos realizó la tarea tres veces, primero con 0 gramos de alcohol, segundo con 0,5 gramos y, por último, con 1 gramo de alcohol en sangre. **Fichero "MR.sav"**.

2.2. Editor de datos, instrucciones y tipo de análisis

Una vez se han introducido los datos en el Editor de datos, antes de ejecutar el procedimiento para realizar el ANOVA, es conveniente comprobar si los datos que tenemos cumplen los supuestos de Normalidad y Homocedasticidad (el procedimiento ANOVA para medidas repetidas del SPSS no tiene la opción de comprobar este supuesto, sólo lo hace para ANOVA con grupos aleatorios).

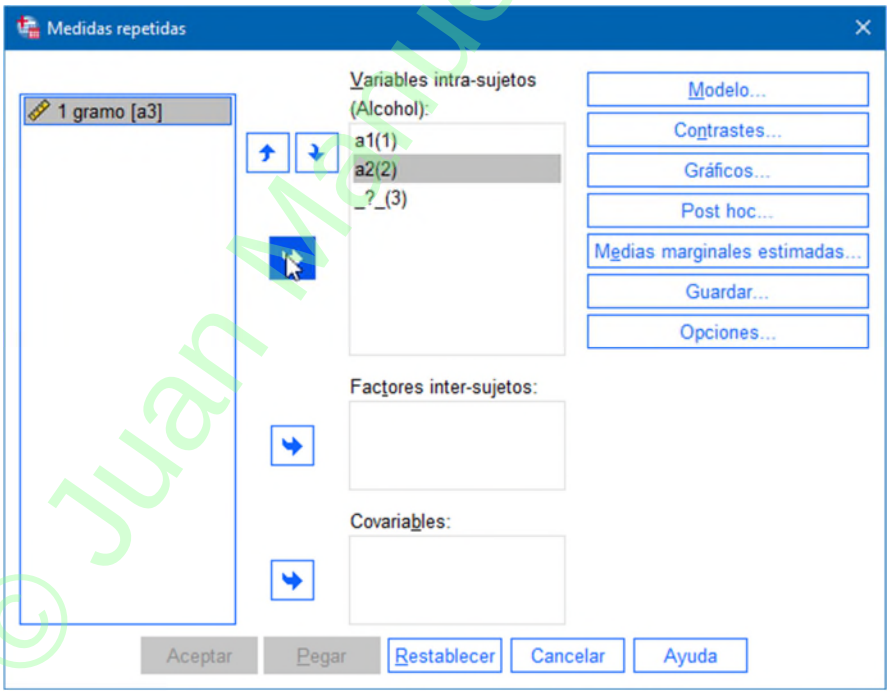
Estos dos supuestos se pueden comprobar mediante el procedimiento *Analizar>Estadísticos descriptivos>Explorar*, introduciendo todas las medidas repetidas (a1, a2 y a3) como "**Dependientes**". Seleccionar "Descriptivos" en la opción "Estadísticos" y en la opción "Gráficos", "gráficos con pruebas de normalidad". En la salida del SPSS obtendremos las varianzas, así como las pruebas de normalidad para las puntuaciones bajo cada nivel de la VI. De las dos pruebas, se suele utilizar la Shapiro-Wilk que es menos conservadora. Se puede asumir que existe homogeneidad de varianzas intra-tratamiento cuando la razón entre la varianza mayor y la menor es menor o igual a 10.

Para ejecutar el ANOVA, se selecciona *Analizar>Modelo Lineal General>Medidas Repetidas*.



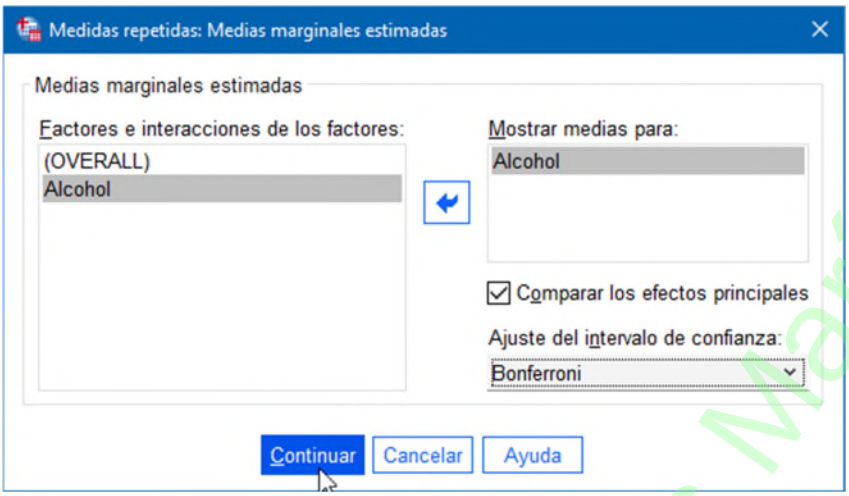
Aparece la ventana Definición de factor(es) de medidas repetidas, en el que hay que introducir el nombre del factor o VI con medidas repetidas (por defecto aparece "factor 1", lo cambiamos por "**Alcohol**"), así como el número de niveles que tiene, en nuestro caso 3. Pinchamos Añadir y después Definir.

Aparecerá entonces la ventana Medidas Repetidas, en la que es necesario definir o especificar los niveles de la VI que se desean contrastar. Para ello, se marca el primer nivel de la ventana de la izquierda, y se pasa a la ventana de la derecha, a continuación, se hace lo mismo con el segundo y con el tercero:

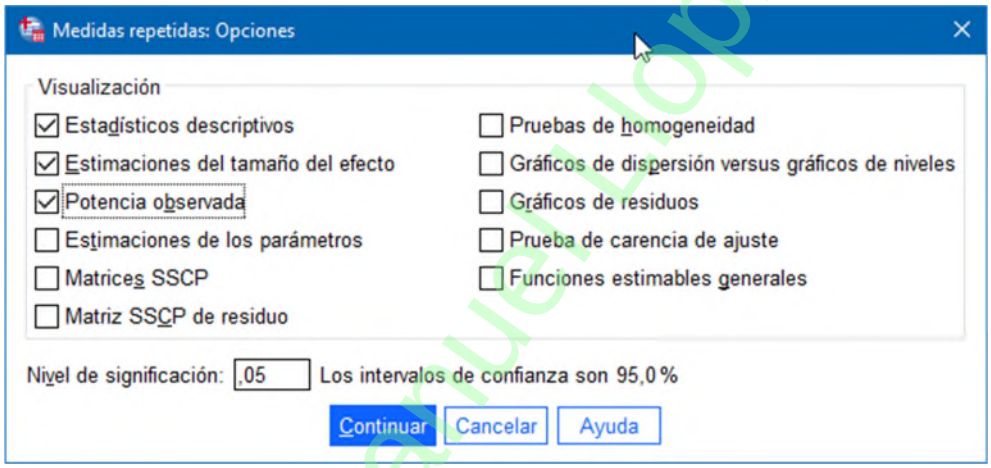


Seleccionamos **Medias marginales estimadas**, y en la ventana que aparece, pinchamos sobre la VI "Alcohol" y la pasamos a la ventana de la derecha ("Mostrar las medias para:") para poder pedir el análisis de factores principales y ajustar mediante la corrección de Bonferroni. Como en este estudio sólo tenemos una VI, el programa comparará todos los niveles de esa VI entre ellos. Recordemos que este

análisis solamente tiene sentido interpretarlo si el ANOVA previo indica que existen diferencias estadísticamente entre los niveles considerados.



Pulsamos en **Opciones** y marcamos las casillas como en la siguiente ventana:



2.3.- Resultados

En primer lugar, la salida proporciona los estadísticos descriptivos.

Estadísticos descriptivos

	Media	Desv. estándar	N
0 gramos	8,17	1,090	24
0,5 gramos	4,50	0,978	24
1 gramo	4,67	1,736	24

A continuación, aparece la tabla resumen para el Análisis de Varianza Multivariado (MANOVA). Después, el resultado de la prueba de esfericidad de Mauchly y la tabla resumen del ANOVA univariado.

Pruebas multivariante[a]

Efecto		Valor	F	gl de hipótesis	gl de error	Sig.	Eta parcial al cuadrado	Parámetro de no centralidad	Potencia observada[c]
Alcohol	Traza de Pillai	0,816	48,868[b]	2,000	22,000	<0,001	0,816	97,735	1,000
	Lambda de Wilks	0,184	48,868[b]	2,000	22,000	<0,001	0,816	97,735	1,000
	Traza de Hotelling	4,443	48,868[b]	2,000	22,000	<0,001	0,816	97,735	1,000
	Raíz mayor de Roy	4,443	48,868[b]	2,000	22,000	<0,001	0,816	97,735	1,000

a. Diseño : Intersección
Diseño intra-sujetos: Alcohol

b. Estadístico exacto

c. Se ha calculado utilizando alpha = ,05

Prueba de esfericidad de Mauchly[a]

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Aprox. Chi-cuadrado	gl	Sig.	Épsilon[b] Greenhouse-Geisser	Huynh-Feldt	Límite inferior
Alcohol	0,989	0,244	2	0,885	0,989	1,000	0,500

Prueba la hipótesis nula de que la matriz de covarianzas de error de las variables dependientes con transformación ortonormalizada es proporcional a una matriz de identidad.

a. Diseño : Intersección
Diseño intra-sujetos: Alcohol

b. Se puede utilizar para ajustar los grados de libertad para las pruebas promedio de significación. Las pruebas corregidas se visualizan en la tabla de pruebas de efectos intra-sujetos.

Pruebas de efectos intra-sujetos

Medida: MEASURE_1

Origen		Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado	Parámetro de no centralidad	Potencia observada[a]
Alcohol	Esfericidad asumida	205,778	2	102,889	50,952	<0,001	0,689	101,904	1,000
	Greenhouse-Geisser	205,778	1,978	104,022	50,952	<0,001	0,689	100,794	1,000
	Huynh-Feldt	205,778	2,000	102,889	50,952	<0,001	0,689	101,904	1,000
	Límite inferior	205,778	1,000	205,778	50,952	<0,001	0,689	50,952	1,000
Error(Alcohol)	Esfericidad asumida	92,889	46	2,019					
	Greenhouse-Geisser	92,889	45,499	2,042					
	Huynh-Feldt	92,889	46,000	2,019					
	Límite inferior	92,889	23,000	4,039					

a. Se ha calculado utilizando alpha = ,05

El resultado de la prueba de Mauchly indica que no se incumple el supuesto de esfericidad, por tanto, es mejor (más potente) el ANOVA (univariado) que el MANOVA (multivariado), y dentro de esta segunda tabla, los resultados para las sumas de cuadrados, grados de libertad, etc. adecuados son los que corresponden a "Esfericidad asumida". El resultado del ANOVA, por tanto, indica que existen diferencias estadísticamente significativas en los aciertos obtenidos entre los tres niveles de alcoholemia. La respuesta a la pregunta ¿entre qué niveles concretamente? nos la proporciona el análisis de factores principales que solicitamos en Opciones y que aparece en la tabla Comparaciones por parejas.

Comparaciones por parejas

Medida: MEASURE_1

(I) Alcohol	(J) Alcohol	Diferencia de medias (I-J)	Desv. Error	Sig.[b]	95% de intervalo de confianza para diferencia[b]	
					Límite inferior	Límite superior
1	2	3,667*	0,393	<0,001	2,651	4,682
	3	3,500*	0,430	<0,001	2,390	4,610
2	1	-3,667*	0,393	<0,001	-4,682	-2,651
	3	-0,167	0,407	1,000	-1,217	0,884
3	1	-3,500*	0,430	<0,001	-4,610	-2,390
	2	0,167	0,407	1,000	-0,884	1,217

Se basa en medias marginales estimadas

*. La diferencia de medias es significativa en el nivel ,05.

b. Ajuste para varias comparaciones: Bonferroni.

Como se puede ver en la tabla, las diferencias estadísticamente significativas en el número de aciertos se encuentran entre el primer nivel (0 gramos de alcohol) y el segundo (0,5 gramos de alcohol), así como entre el primero y el tercero (1 gramo de alcohol). Remitiéndonos a las puntuaciones medias de cada nivel, vemos que el mayor número de aciertos corresponde a 0 gramos de alcohol en sangre. Sin embargo, no se han encontrado diferencias estadísticamente significativas entre el nivel segundo y el tercero. Por tanto, no podemos afirmar, con estos datos, que la disminución de 0,5 gramos de alcohol en sangre, respecto de 1 gramo, produzca mejoras en el número de aciertos que obtienen los sujetos en la tarea de simulación de conducción de automóviles.

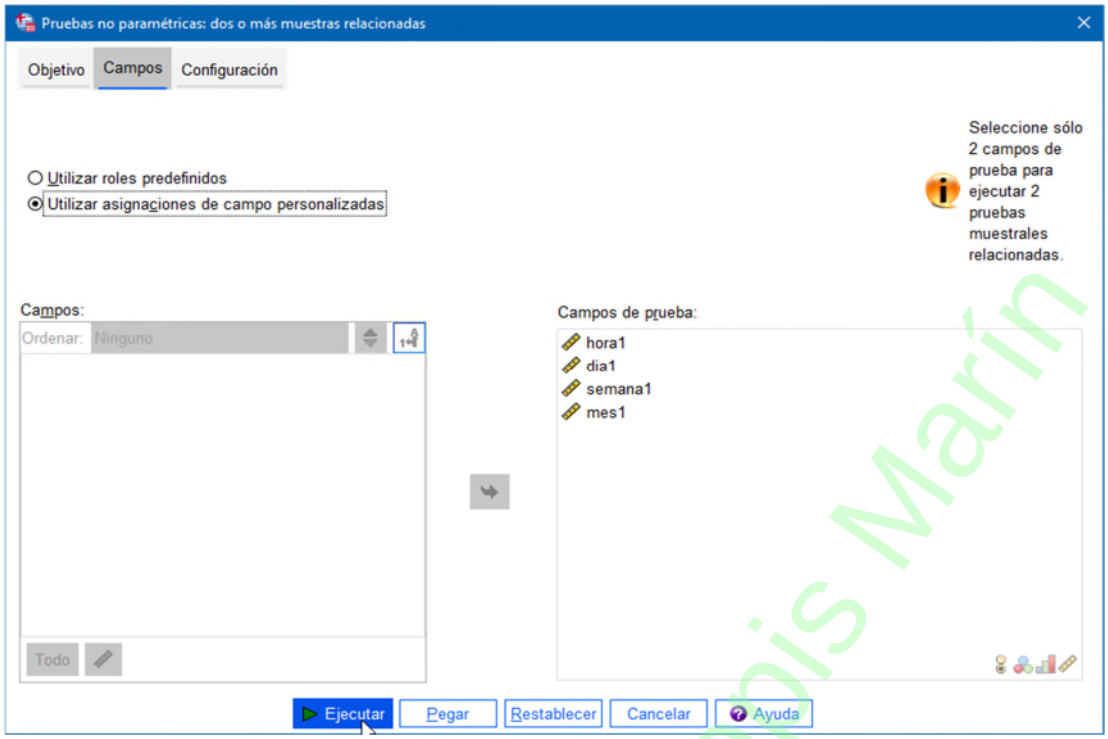
3.- Alternativas para datos categóricos

3.1. Contrastes para más de dos muestras relacionadas (mínimo datos ordinales)

Es similar al ANOVA de un factor con medidas repetidas o intrasujetos. Las ventajas fundamentales respecto al ANOVA es que no necesita establecer supuestos sobre las poblaciones (normalidad y homocedasticidad) y permite trabajar, al menos, con datos ordinales. La prueba más utilizada es la de Friedman, aunque el SPSS nos ofrece una prueba adicional que nos lleva a conclusiones similares, esta es la W de Kendall (coeficiente de concordancia). Veamos un ejemplo:

En investigaciones sobre la memoria se ha intentado estudiar el efecto distorsionante del paso del tiempo. En un estudio, a un grupo de 9 sujetos se les presentó una historia escrita que debían memorizar durante 20 minutos. Tras distintos momentos se les dijo que escribieran en un papel la historia, siendo esta evaluada por un grupo de expertos. Los momentos de medida se sucedieron a una hora, un día, una semana y un mes. ¿Podemos afirmar que el paso del tiempo influye sobre la calidad del recuerdo? El nivel de significación se establece en 0,05. **Fichero: "Friedman.sav"**

Para responder a esta pregunta con ayuda del SPSS; Analizar>Pruebas No paramétricas>Muestras relacionadas. De las tres pestañas que aparecen, en la pestaña Campos, colocamos todas las variables en campos de prueba y pulsamos ejecutar, el programa detectará nuestros datos y aplicará el análisis adecuado (en este caso la prueba de Friedman). Si queremos un análisis personalizado, lo elegimos en las pestañas Objetivo y Configuración (seleccionando la prueba de Friedman).



La salida básica es:

Pruebas no paramétricas

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.[a,b]	Decisión
1	Las distribuciones de hora1, dia1, semana1 y mes1 son iguales.	Prueba de Friedman para muestras relacionadas para análisis de la varianza de dos factores por rangos	<0,001	Rechace la hipótesis nula.

a. El nivel de significación es de 0,050.

b. Se muestra la significancia asintótica.

Prueba de Friedman para muestras relacionadas para análisis de la varianza de dos factores por rangos

hora1, dia1, semana1, mes1

Resumen de la prueba de Friedman para muestras relacionadas para análisis de la varianza de dos factores por rangos

N total	9
Estadístico de prueba	18,556
Grado de libertad	3
Sig. asintótica (prueba bilateral)	<0,001

Lo que nos lleva a concluir que existen diferencias entre las medidas consideradas. Ahora debemos explorar entre qué grupos se encuentran esas diferencias. El resultado de la comparación aparece en la siguiente tabla:

Comparaciones por parejas

Sample 1-Sample 2	Estadístico de prueba	Error estándar	Estadístico de prueba estándar	Sig.	Sig. ajust.[a]
semana1-mes1	-0,278	0,609	-0,456	0,648	1,000
semana1-dia1	0,778	0,609	1,278	0,201	1,000
semana1-hora1	2,278	0,609	3,743	<0,001	0,001
mes1-dia1	0,500	0,609	0,822	0,411	1,000
mes1-hora1	2,000	0,609	3,286	0,001	0,006
dia1-hora1	1,500	0,609	2,465	0,014	0,082

Cada fila prueba la hipótesis nula que las distribuciones de la Muestra 1 y la Muestra 2 son iguales.
Se muestran las significaciones asintóticas (pruebas bilaterales). El nivel de significación es de 0,050.

a. Los valores de significación se han ajustado mediante la corrección Bonferroni para varias pruebas.

Obtenemos que existen diferencias entre una hora y una semana y entre una hora y un mes, pero no en el resto de las comparaciones (se mira la columna de significación ajustada). Viendo las medias que aparecen en los gráficos posteriores a la tabla de comparaciones por parejas, hay significativamente más recuerdo al cabo de una hora que al cabo de una semana o de un mes.

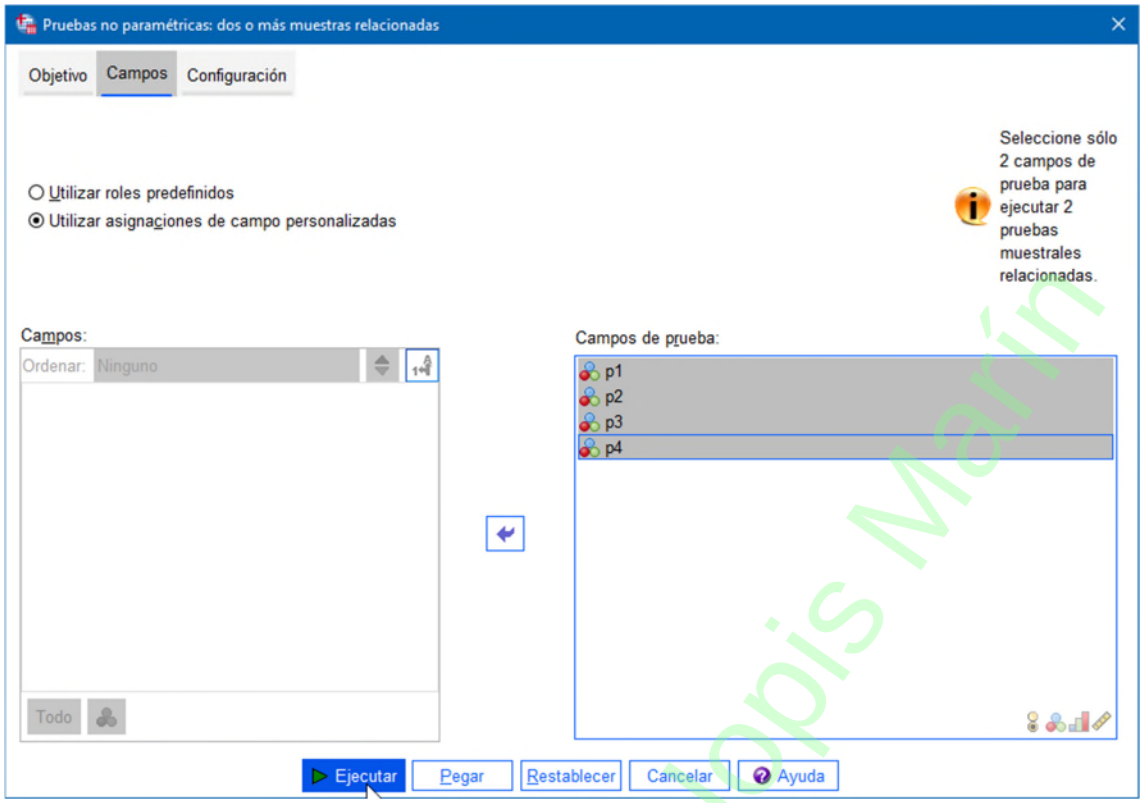
3.2. Contrastes para más de dos muestras relacionadas (Datos nominales)

Nos permite contrastar hipótesis sobre la igualdad de más de dos proporciones para muestras dependientes con medidas cualitativas o nominales donde se suelen utilizar como medidas las proporciones. La prueba más utilizada es la Q de Cochran y se suele aplicar a diseños de medidas repetidas donde la VD o variable de respuesta sólo puede tomar dos valores (variable dicotómica).

Ejemplo: Un psicólogo quiere averiguar si las 4 preguntas de un test que ha construido tienen o no la misma dificultad. Para ello, selecciona aleatoriamente una muestra de 10 sujetos que responden a las 4 preguntas. ¿Se puede afirmar que las preguntas difieren en dificultad? El nivel de significación se establece en 0,05. **Fichero: "QCochran.sav".**

Para responder a esta pregunta con ayuda del SPSS; *Analizar>Pruebas No paramétricas>Muestras relacionadas*. De las tres pestañas que aparecen, en la pestaña Campos, colocamos todas las variables en campos de prueba y pulsamos ejecutar, el programa detectará nuestros datos y aplicará el análisis adecuado (en este caso la prueba Q de Cochran). Si queremos un análisis personalizado, lo elegimos en las pestañas Objetivo y Configuración (seleccionando la prueba Q de Cochran).

Introducimos las diferentes medidas en contrastar variables y pulsamos Q de Cochran.



La salida básica es:

Pruebas no paramétricas

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.[a,b]	Decisión
1	Las distribuciones de p1, p2, p3 y p4 son iguales.	Prueba Q de Cochran para muestras relacionadas	0,019	Rechace la hipótesis nula.

a. El nivel de significación es de 0,050.

b. Se muestra la significancia asintótica.

Prueba Q de Cochran para muestras relacionadas

p1, p2, p3, p4

Resumen de prueba Q de Cochran de muestras relacionadas

N total	10
Estadístico de prueba	10,000
Grado de libertad	3
Sig. asintótica (prueba bilateral)	0,019

Lo que nos lleva a concluir que existen diferencias entre las medidas consideradas. Ahora debemos explorar entre qué grupos se encuentran esas diferencias. El resultado de la comparación aparece en la siguiente tabla:

Comparaciones por parejas

Sample 1-Sample 2	Estadístico de prueba	Error estándar	Estadístico de prueba estándar	Sig.	Sig. ajust.[a]
p4-p2	0,300	0,224	1,342	0,180	1,000
p4-p1	0,400	0,224	1,789	0,074	0,442
p4-p3	0,700	0,224	3,130	0,002	0,010
p2-p1	0,100	0,224	0,447	0,655	1,000
p2-p3	-0,400	0,224	-1,789	0,074	0,442
p1-p3	-0,300	0,224	-1,342	0,180	1,000

Cada fila prueba la hipótesis nula que las distribuciones de la Muestra 1 y la Muestra 2 son iguales.
Se muestran las significaciones asintóticas (pruebas bilaterales). El nivel de significación es de 0,050.

a. Los valores de significación se han ajustado mediante la corrección Bonferroni para varias pruebas.

Podemos ver que existen diferencias en dificultad entre las preguntas 3 y 4 (significación ajustada menor que 0,05).

El gráfico de barras apiladas nos indica que existen, significativamente, más errores en la pregunta 4 que en la 3, no habiendo diferencias significativas entre el resto de las comparaciones: