2017

Manual de Estadística Avanzada en Medicina

SERGIO BARROSO HERNÁNDEZ

Aulia. Centro de estudios de Badajoz

MANUAL DE ESTADÍSTICA AVANZADA EN MEDICINA

MANUAL DE ESTADÍSTICA AVANZADA EN MEDICINA

SERGIO BARROSO HERNÁNDEZ

UNIVERSIDAD DE EXTREMADURA

Cáceres
2018

Edita:
Universidad de Extremadura. Servicio de Publicaciones.
C/ Caldereros, 2. Planta 2ª
10071 Cáceres
Telf. 927 257 041 • Fax 927 257 046
e-mail: publicac@unex.es
http://www.unex.es/publicaciones

ISBN: 978-84-697-4485-7
Depósito Legal: BA-187/2018

Impreso en España - Printed in Spain

Impresión:
Tecnigraf, S.A.
Teléfono 924 286 006
www.tecnigraf.com

A mi mujer, Elena,
y a mis hijas,
Carmen y María.

ÍNDICE

PRÓLOGO

Durante el presente curso vamos a aprender los conocimientos necesarios para llevar a cabo la realización de la mayoría de los trabajos de investigación que vamos a realizar a lo largo de toda nuestra carrera profesional. La investigación en el campo de la Medicina, no es algo obligatorio y en la mayoría de las ocasiones nos va a suponer una dedicación de un tiempo importante, generalmente fuera de nuestro horario laboral con consumo de nuestro tiempo libre; sin embargo, supone un importante estímulo y nos puede abrir la mente a la hora de enfrentar problemas de nuestra práctica clínica diaria.

Para llevar a cabo un estudio de investigación clínica vamos a necesitar:

- Saber elegir el tipo de estudio adecuado a lo que queremos realizar.
- Tener los conocimientos estadísticos necesarios para llevar a cabo el estudio.
- Saber utilizar un programa estadístico para manejar, analizar y representar los datos del estudio.
- Y fundamentalmente, tiempo.

El objetivo del presente manual es darle solución a los tres primeros puntos anteriores. Desgraciadamente el factor tiempo es una variable no modificable y va a depender del propio investigador; pero debemos tener en cuenta que llevar a cabo un estudio de investigación es un proceso laborioso y que consume mucho tiempo, por lo que si no disponemos de él, es preferible no iniciarlo puesto que los trabajos hechos con prisas suelen presentar gran cantidad de imperfecciones que pueden invalidar los resultados y al final habremos perdido el poco tiempo que le hayamos dedicado para nada.

Para ello vamos en primer lugar a conocer las principales características del programa estadístico más utilizado que es SPSS. Aprenderemos a familiarizarnos con el programa y a adquirir un manejo fluido del mismo a lo largo del temario. Debemos saber que SPSS "no trabaja solo", es el investigador el que tiene que guiarlo. No es una coctelera en la que se metan datos y salgan resultados. Veremos a lo largo del curso, que según los objetivos del estudio debemos indicarle a SPSS que haga una cosa u otra.

Posteriormente se detallarán las principales herramientas estadísticas para analizar cualquier estudio, aumentando la complejidad de las mismas a medida que vayamos avanzando en el curso.

El manual está pensado para que sea lo más práctico posible, sin laboriosas y complicadas explicaciones, detallando paso a paso cómo obtener los resultados de los ejemplos. Debe servir como manual de consulta para cuando queramos hacer cualquier estudio, pues lógicamente las cosas se olvidan y vamos a necesitar un documento recordatorio con los pasos necesarios en cada caso.

PROCESADO DE DATOS CON SPSS

INTRODUCCIÓN

SPSS es uno de los procesadores de datos estadísticos más utilizado en el mundo, pero hay otros como STATA, SAS, etc. Vamos a utilizar SPSS por su sencillez, cuenta con una interfaz intuitiva y con una hoja de sintaxis que nos informa de nuestros errores. La principal desventaja es su alto precio, aunque podemos disponer de versiones de prueba gratuitas por un periodo de 15 días. En este curso vamos a utilizar SPSS versión 23.

En primer lugar debemos tener en cuenta que SPSS no es una base de datos. Tampoco lo es Excel (es una hoja de cálculo). Como ejemplos de bases de datos tenemos FileMaker o Access. Las bases de datos están diseñadas para introducir y almacenar datos de una manera sencilla y ordenada permitiendo la posibilidad de crear prevalidaciones o postvalidaciones que impidan la introducción de datos erróneos. SPSS no permite esta posibilidad. Por ejemplo, si estamos recogiendo la variable "creatinina", podemos introducir el valor 4.8 ó 48 en SPSS y no alertarnos de ningún tipo de error. Sin embargo en las bases de datos podemos crear criterios para impedirlo. Cuando la base de datos cuenta con pocas variables y pocos casos, es fácil darse cuenta de dónde está el error, pero cuando manejamos muchas variables con muchos casos, esta labor se complica y la introducción de múltiples valores erróneos puede artefactarnos los resultados.

Si el estudio que queremos hacer cuenta con pocas variables o pocos casos, Excel por su sencillez a la hora de introducir valores, es una opción válida. Pero cuando pensemos hacer un estudio con muchas variables o muchos casos, debemos pensar en usar o diseñar nuestra propia base de datos. En ningún caso es recomendable introducir datos sobre SPSS.

El uso de los distintos estadísticos necesarios para el análisis de los datos se verá a medida que los vayamos estudiando en los distintos temas. Pero previo a ello debemos tener correctamente bien definidas las variables existentes y creadas aquellas que vayamos a necesitar. Este va a ser el objetivo de esta primera parte del curso.

Para nombrar las variables de nuestro estudio, debemos hacerlo de manera sencilla, con nombres no excesivamente largos. El nombre nos debe permitir identificar la variable con facilidad. Por ejemplo para la variable "Fecha de Nacimiento" podemos llamarla "FNac". Hay variables que se repiten en casi todos los estudios, de manera que si las

nombramos siempre igual podremos utilizar sintaxis guardadas como luego veremos. En ningún caso el nombre de una variable debe empezar por un número, pues SPSS no la va a reconocer.

SPSS nos permite crear variables a partir de las existentes, y realizar cálculos, por lo tanto no tiene sentido que perdamos el tiempo a la hora de introducir los datos derivados de transformaciones matemáticas en nuestra base de datos. Por ejemplo calcular la edad (bastará introducir la fecha de nacimiento y la fecha actual), o el IMC (bastará introducir el peso y la talla), etc.

Todas las acciones que vayamos haciendo sobre nuestra base de datos, vamos a ir pegándolas a una hoja de **sintaxis**. Esta hoja va a ser nuestro diario. La mayoría de las personas no iniciadas en el uso del SPSS desconocen su existencia. Tiene la ventaja de recoger todas nuestras actuaciones, las cuales podrán ser utilizadas en un futuro si por ejemplo ampliamos el tamaño de nuestra muestra (no tenemos que empezar otra vez de cero, tan sólo volver a ejecutar la sintaxis que tenemos guardada); las acciones guardadas pueden ser utilizadas para otra base de datos distinta; por ejemplo, si creamos las ecuaciones para estimar el filtrado glomerular mediante la fórmula CKD- EPI –la cual sea dicho de paso es tremendamente engorrosa de crear- para una base de datos, podremos volver a utilizarla para otros estudios distintos donde se necesite sin necesidad de volver a crearla; nos permite modificar algunos parámetros a nuestro antojo; podemos escribir sobre ella anotaciones que nos orienten sobre lo que se está haciendo para cuando queramos revisarlo (para escribir sobre la sintaxis una anotación debemos empezar la frase con un asterisco " * " y terminarla con un punto " . "; la frase en cuestión se volverá gris y de esta manera SPSS sabe que eso no es un comando ejecutable); y por último, tenemos que recordar que cuando realizamos un estudio de investigación y exponemos unos resultados, alguien (como una auditoría de una revista) nos puede pedir que demostremos de qué manera se han obtenido los resultados del estudio en concreto y demostrar que no han sido manipulados.

Para este curso vamos a utilizar principalmente la base de datos "acidosis.accdb". Es una base de datos en formato Access con valores reales de un estudio. Esta base de datos cuenta con restricciones de tal manera que se imposibilita la introducción de valores erróneos casi con un 100% de seguridad.

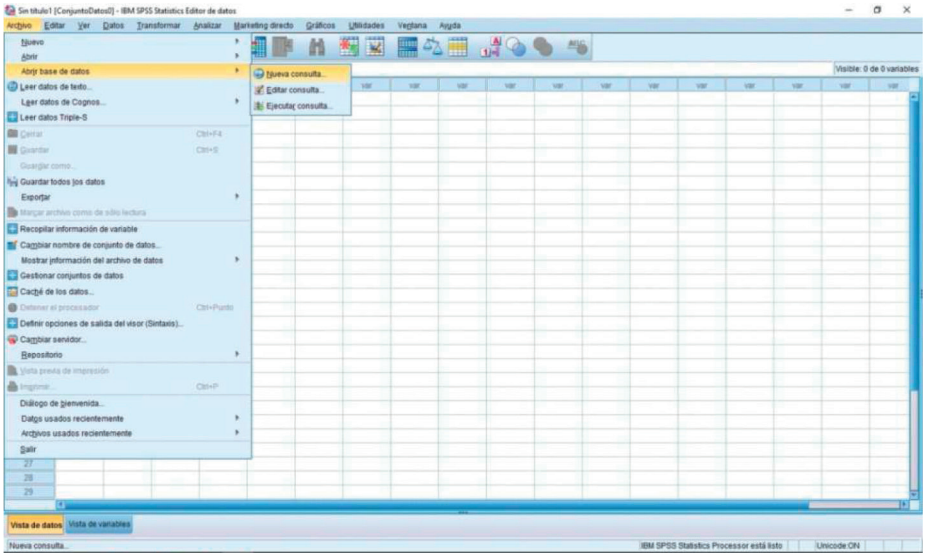
IMPORTACIÓN DE DATOS

Para la importación de datos con SPSS primero debemos abrir la aplicación del programa estadístico y seleccionar la acción:

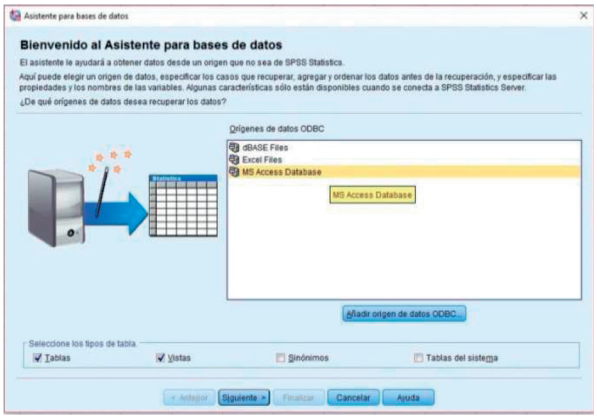
Abrir base de datos → Nueva consulta.

Otra opción disponible es:

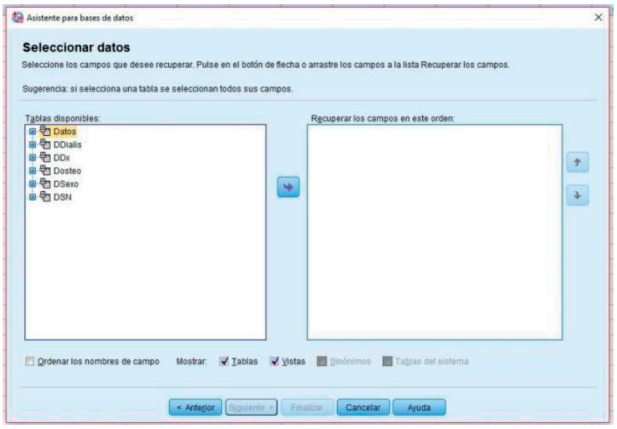
Abrir → Datos, y en el menú de diálogo "Archivos de tipo" seleccionar el tipo de archivo que vayamos a abrir.



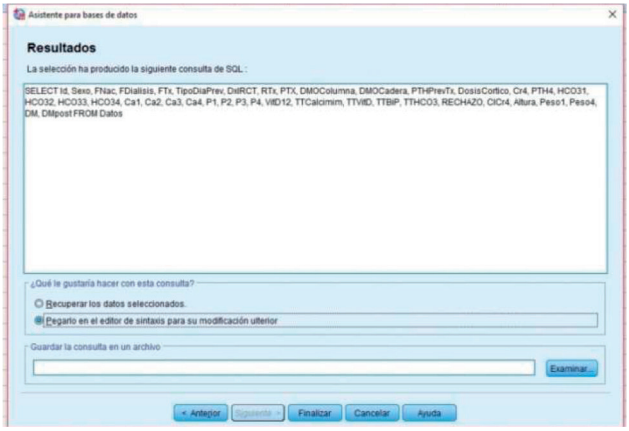
El siguiente paso es seleccionar el tipo de archivo, en este caso "MS Access Database". Seleccionaremos la opción "Vistas" para poder ver todas las extensiones del archivo.



En el siguiente paso aparecen todos los elementos de los que dispone la base de datos. Seleccionaremos la tabla que contenga los datos de nuestro estudio y los arrastraremos hacia el cuadro de la derecha.

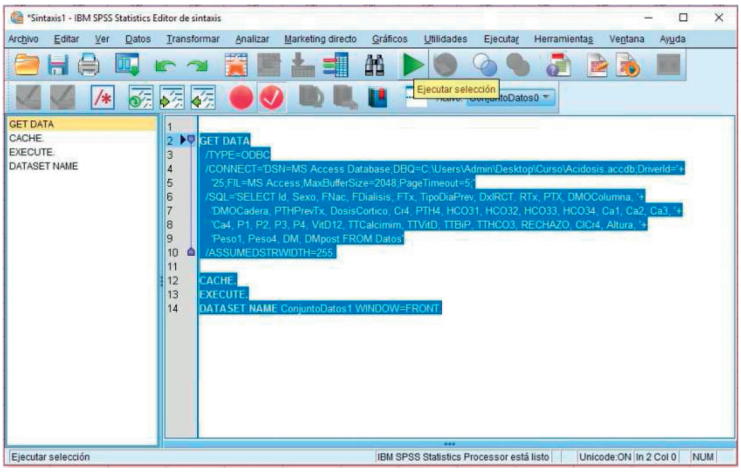


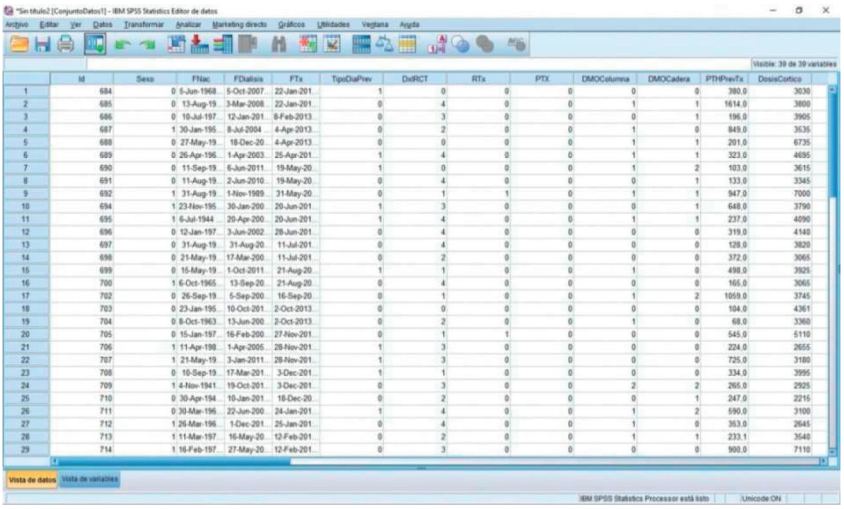
Haciendo Clic sobre el botón "Siguiente" llegaremos al último cuadro de diálogo donde seleccionaremos la opción "Pegarlo en el editor de sintaxis…" y haremos Clic sobre el botón "Finalizar".



Se nos abrirá una nueva hoja que será nuestra hoja de sintaxis sobre la que iremos guardando todos los procesos que vayamos realizando. Seleccionamos todo el texto y hacemos Clic sobre la flecha verde de la cinta de opciones de arriba para "Ejecutar la selección". A continuación se nos abrirá la hoja de datos de SPSS con todas las variables y valores de nuestro estudio, las cuales procederemos a continuación a definir. Guardaremos previamente la hoja de datos y la hoja de sintaxis en nuestro ordenador; en este caso las guardaremos con el nombre "Acidosis" creándose los archivos "Acidosis.sav" (hoja de datos) y "Acidosis.sps" (hoja de sintaxis) en la ubicación que hayamos elegido.

Cuando vamos a guardar la hoja de datos, SPSS nos da la opción de guardar sólo aquellas variables que nos parezcan oportunas con la opción "Variables", por ejemplo si queremos trabajar con 2 hojas de datos con distintas variables.

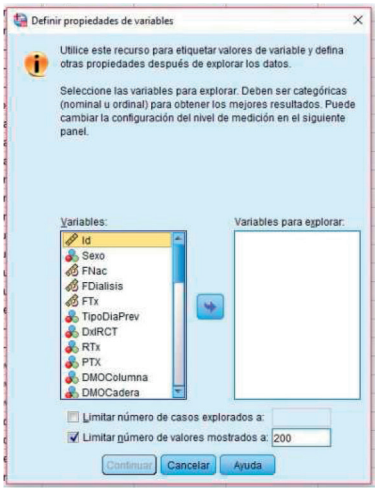




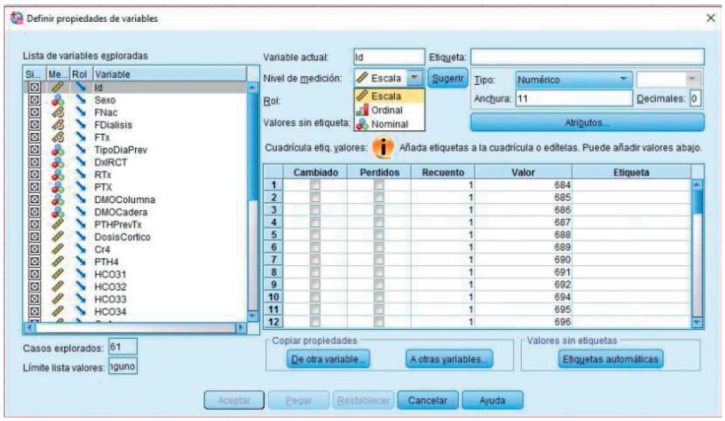
DEFINICIÓN DE LAS VARIABLES

Para que SPSS funcione y analice correctamente los datos, las variables deben estar perfectamente bien definidas. Para ello seleccionamos la opción de la cinta de opciones superior:

Datos → Definir propiedades de variables; y seleccionamos aquellas que vamos a definir.



Seleccionamos las variables que queramos hacia el cuadro de la derecha. En el siguiente cuadro de diálogo podemos escribir el nombre de la variable en el cuadro "Etiqueta"; lo que pongamos aquí es lo que va a aparecer después cuando hagamos gráficos, tablas, etc. Por tanto es recomendable no cometer errores ortográficos. Se acostumbra a poner el nombre de la variable y la medida en la que se ha realizado (por ejemplo para la variable Cr se pondría "Creatinina plasmática (mg/dl)").



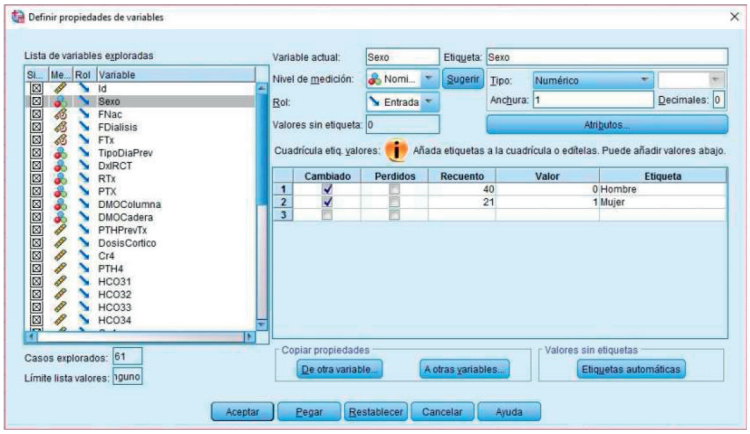
Existen tres opciones de variables ("Escala", "Ordinal" y "Nominal"). Cada una hace referencia a un tipo distinto de variables, y su definición correcta es muy importante para que SPSS sepa de qué tipo de variables se trata para el análisis. Como ejemplo de "Escala" tenemos la creatinina plasmática; como "Ordinal" se consideran aquellas variables cualitativas que siguen un orden el cual implican una consideración distinta, por ejemplo los estadios de insuficiencia renal crónica (IRC), tenemos estadios del 1 al 5 siguiendo un pronóstico diferente a medida que aumenta el estadio. Como "Nominal" son aquellas variables cualitativas que no siguen ningún orden, por ejemplo la variable Sexo, o la variable Diagnóstico. En el cuadro "Tipo" nos indicará de qué tipo de variable se trata: Numérica, Fecha, Cadena…

Podemos modificar en qué formato aparecerán las variables numéricas y fecha. Para las numéricas, podemos definir el número de decimales que aparecen a la derecha de la coma, y la anchura total del valor. Primero se modifica el número de decimales y posteriormente la anchura del valor teniendo en cuenta que la coma y el signo negativo (si lo hubiera) también se contabilizan. El valor con mayor anchura conseguiremos verlo moviendo la barra de dirección vertical de la derecha para ver el último valor que sería el valor más ancho con valor positivo, o viendo el primer valor que sería el valor más ancho si tiene valor negativo. Para modificar el formato de una fecha, se selecciona el que deseemos del cuadro de diálogos desplegable a la derecha del campo "Fecha".

Para las variables categóricas podemos definir las etiquetas de cada categoría en el campo "Etiqueta" que hay al lado del "Valor". Lo que escribamos aquí también va a aparecer en las tablas, gráficos, etc., y por tanto es recomendable escribirlo sin errores.

Si consideramos que algún valor es erróneo, o aún no siendo erróneo, no queremos que SPSS lo tenga en cuenta a la hora del análisis (por ejemplo porque sea un valor muy extremo), podemos anularlo seleccionando el cuadro de la opción "Perdidos". SPSS lo considerará como valor "perdido por el usuario" (USER MISSING).

Ejemplos de algunas variables de nuestro estudio:



Tras haber definido todas las variables, haremos Clic en la opción "Pegar" para registrarlo en nuestra hoja de sintaxis. Seleccionaremos todo (lo que hemos hecho nuevo) y haremos Clic en la flecha verde para "Ejecutar la selección".

CREACIÓN DE NUEVAS VARIABLES (COMPUTE, IF)

La creación de variables es uno de los procedimientos más usados e importantes de un estudio. En cualquier estudio que vayamos a realizar, vamos a recoger las variables que estimemos importantes, pero posteriormente será necesaria la creación de otras nuevas a partir de las existentes. Por ejemplo, calcular la edad de los pacientes, creación de puntos de corte, cálculos matemáticos…

SPSS dispone de infinidad de opciones para creación de nuevas variables. Con la ayuda del programa y nuestra imaginación podemos crear cualquier variable que se nos ocurra. El proceso se inicia haciendo Clic sobre la opción de la cinta de opciones:

Transformar → Calcular variable…; Tras ello aparece el cuadro de diálogo siguiente:

En el recuadro superior izquierdo "Variable objetivo" escribiremos de manera simplificada el nombre que vayamos a darle a la nueva variable. Seleccionamos la(s) variable(s) a partir de la cual(es) haremos los cálculos y los desplazaremos hacia el recuadro de la derecha "Expresión numérica". Aquí podremos hacer infinidad de cálculos como si fuera una calculadora o bien con las opciones del recuadro "Grupo de funciones".

Vamos a realizar los cálculos más frecuentes en cualquier estudio.

Cálculo de la edad

Generalmente se acostumbra a recoger en los datos la fecha de nacimiento, que junto con la fecha que nosotros seleccionemos, calcularemos la edad de los pacientes. No es correcto recoger directamente la edad de los pacientes en nuestra base de datos puesto que es la edad en "años cumplidos"; si lo hacemos, daremos por hecho que una persona que nazca el 1 de enero y otra el 31 de diciembre van a tener la misma edad, sin embargo es obvio que esto no es correcto y este error será importante sobre todo en aquellas patologías con rápida evolución, por ejemplo en meses. Si se recoge la edad en "años cumplidos" debemos sumar 0.5 a todos los valores; de esta manera compensaremos el error de los valores extremos (cercanos a enero o a diciembre).

Para el cálculo de la edad al momento del trasplante en nuestra base de datos seleccionamos la variable "fecha del trasplante" menos "fecha de nacimiento". Posteriormente seleccionamos la opción "extracción de la duración del tiempo" de las opciones disponibles en el cuadro "Grupo de funciones", tal como se muestra en la siguiente imagen. El valor resultante será en días. Para transformarlo en años basta con dividirlo por 365.25 (es el valor que tiene en cuenta los años bisiestos). Si lo que queremos es transformarlo en meses dividiremos por 30.4375. Y si lo queremos en semanas lo dividimos entre 52.1786. Las sintaxis serían las siguientes:

- COMPUTE Edad=CTIME.DAYS(FTx - FNac)/365.25 (en años exactos).
- COMPUTE Edad=CTIME.DAYS(FTx - FNac)/30.4375 (en meses exactos).
- COMPUTE Edad=CTIME.DAYS(FTx - FNac)/52.1786 (en semanas exactas).

Después de la creación de la nueva variable, por supuesto hay que "Definir" la nueva variable como hemos aprendido en el apartado anterior.

Para el cálculo del tiempo en meses que los pacientes llevan en diálisis antes del trasplante, la sintaxis correspondiente sería:

COMPUTE TiempHD=CTIME.DAYS(FTx - FDialisis)/30.4375.

Cálculo con operaciones matemáticas

En esta opción tan sólo debemos escribir la operación matemática que estimemos oportuna en el recuadro "Expresión numérica".

Vamos a calcular la variable Índice de Masa Corporal al momento del trasplante y al año (IMCinic, IMCaño). Para ello debemos seleccionar las variables peso y talla, y escribir la operación matemática correspondiente, tal que así:

*COMPUTE IMCinic=Peso1/(Altura ** 2).*

*COMPUTE IMCaño=Peso4/(Altura ** 2).*

Debemos tener mucho cuidado con los paréntesis. Los dos asteriscos (**) señalan el número al que será elevado el valor (en este caso al cuadrado). Un solo asterisco (*) es el símbolo de multiplicar.

Si lo que queremos es crear una variable binaria a partir de una cuantitativa la forma correcta de hacerlo es través de este procedimiento (no a través de la opción "Recodificar" que más adelante veremos). Por ejemplo, queremos crear la variable "Insuficiencia renal crónica" como aquellos pacientes con un Aclaramiento de creatinina menor de 60 (ClCr <60 ml/min). Para ello tan sólo tendremos que realizar la operación siguiente:

COMPUTE IRC=ClCr4 < 60.

De esta manera a todos los pacientes con valores válidos (no tiene en cuenta los valores perdidos o missing) de ClCr menor que 60 (no incluye el valor 60) les otorga el valor 1, y a todos los demás el valor 0.

Por ejemplo, si queremos dividir a los pacientes en hipercalcémicos o no al año del trasplante considerando hipercalcemia un calcio superior a 10.5 mg/dl, la operación sería la siguiente:

COMPUTE HiperCa=Ca4 > 10.5.

De esta forma sólo se tiene en cuenta los valores válidos de la variable (es decir, sin tener en cuenta los Missing values). Si no se hace de esta manera es posible que a los valores perdidos les otorgue el valor 0 y los englobe en uno de los 2 grupos creados.

Vamos con algo más complicado:

Queremos dividir al grupo de pacientes en "pacientes acidóticos" y "pacientes no acidóticos" durante el primer año. El criterio va a ser que los pacientes tengan al menos las ¾

partes de las analíticas con acidosis (entendida ésta como un bicarbonato <24 mmol/l). Tenemos que proceder de la siguiente manera:

- Dado que tenemos recogidas 4 muestras de bicarbonato, una por cada trimestre del año, vamos a crear la variable acidosis Sí o No para cada trimestre; es decir, queremos construir una variable binaria 0 ó 1 y para esto hemos dicho que lo mejor es hacer lo que se ha explicado en el apartado anterior. Para cada trimestre, la sintaxis sería:

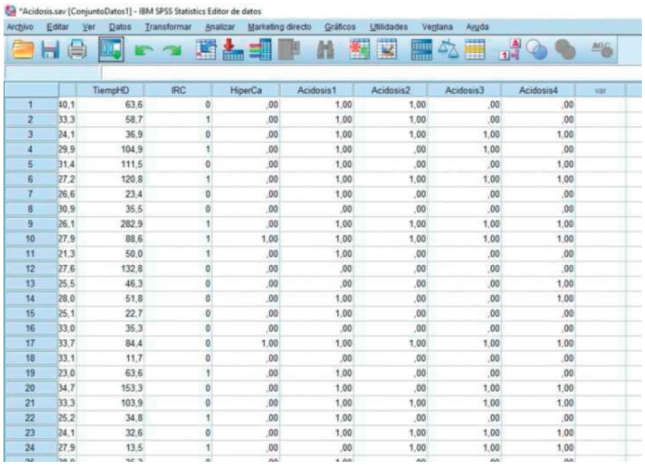
COMPUTE Acidosis1=HCO31 < 24.

COMPUTE Acidosis2=HCO32 < 24.

COMPUTE Acidosis3=HCO33 < 24.

COMPUTE Acidosis4=HCO34 < 24.

EXECUTE.

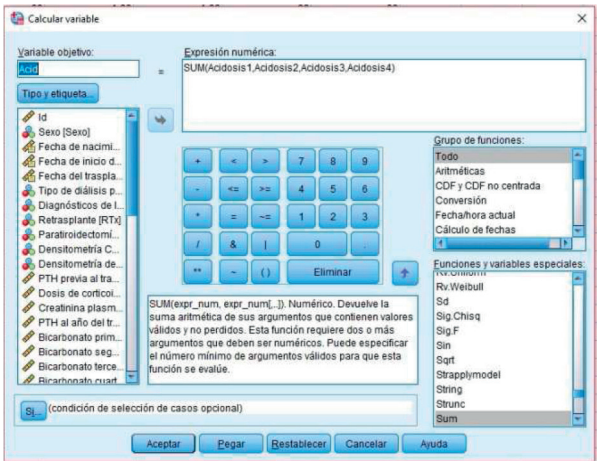


		TiempHD	IRC	HiperCa	Acidosis1	Acidosis2	Acidosis3	Acidosis4	var
1	40,1	63,6	0	,00	1,00	1,00	,00	,00	
2	33,3	58,7	1	,00	1,00	1,00	,00	,00	
3	24,1	36,9	0	,00	1,00	1,00	1,00	1,00	
4	29,9	104,9	1	,00	1,00	,00	1,00	,00	
5	31,4	111,5	0	,00	1,00	,00	,00	1,00	
6	27,2	120,8	1	,00	1,00	1,00	1,00	1,00	
7	26,6	23,4	0	,00	1,00	,00	,00	,00	
8	30,9	35,5	0	,00	,00	,00	,00	,00	
9	26,1	282,9	1	,00	1,00	1,00	1,00	1,00	
10	27,9	88,6	1	1,00	1,00	1,00	1,00	1,00	
11	21,3	50,0	1	,00	1,00	,00	,00	,00	
12	27,6	132,8	0	,00	,00	,00	,00	,00	
13	25,5	46,3	0	,00	,00	,00	,00	1,00	
14	28,0	51,8	0	,00	1,00	,00	,00	1,00	
15	25,1	22,7	0	,00	1,00	,00	,00	,00	
16	33,0	35,3	0	,00	,00	,00	,00	,00	
17	33,7	84,4	0	1,00	1,00	1,00	1,00	1,00	
18	33,1	11,7	0	,00	,00	,00	,00	,00	
19	23,0	63,6	1	,00	1,00	,00	,00	,00	
20	34,7	153,3	0	,00	1,00	,00	1,00	1,00	
21	33,3	103,9	0	,00	1,00	1,00	1,00	1,00	
22	25,2	34,8	1	,00	1,00	,00	,00	,00	
23	24,1	32,6	0	,00	1,00	1,00	1,00	1,00	
24	27,9	13,5	1	,00	,00	1,00	1,00	1,00	

- Ahora debemos crear otra variable que recoja si las ¾ partes de estas analíticas son acidosis o no. Para ello podemos utilizar la opción "SUM" del SPSS que nos sumará el valor de estas variables. El procedimiento se recoge en la siguiente sintaxis e imagen.

COMPUTE Acid=SUM(Acidosis1,Acidosis2,Acidosis3,Acidosis4).

EXECUTE.



De esta forma se nos creará la variable "Acid", que tendrá un valor de 0 si ninguna analítica trimestral es acidótica, 1 si es el 25%, 2 si es el 50%, 3 si es el 75% y 4 si todas son acidóticas.

- Por tanto ahora tan sólo queda volver a crear una variable binaria con un punto de corte en el 3 que divide a los pacientes en aquellos con más o menos del 75% de las analíticas acidóticas. La sintaxis sería:

COMPUTE EstadAcido=Acid >= 3.

EXECUTE.

Ahora quedaría definir las propiedades de la nueva variable creada. Llamarla simplemente "Acidosis".

Para estar seguro de lo que vamos haciendo, y comprobar que las variables que vamos creando están bien, podemos comprobarlo de manera visual. Para ello podemos utilizar la opción "LIST VARIABLES" escribiéndolo en la sintaxis y a continuación las variables que queramos listar. En esta caso las variables del bicarbonato y las nuevas creadas. Se nos abrirá una hoja de resultados con un listado de estas variables donde podremos comprobar que todo está correcto.

Si en lugar de definir a un paciente como "acidótico" cuando ¾ partes de las analíticas del año presentan acidosis, lo definiéramos como aquél paciente que presente la media de bicarbonato durante el año inferior a 24 se nos podría ocurrir sumar los valores de bicarbonato de las 4 variables y dividirlo por 4 según la siguiente sintaxis:

COMPUTE MediaBicar=(HCO31 + HCO32 + HCO33 + HCO34)/4.

EXECUTE.

Este procedimiento sería correcto cuando todos los casos en las cuatro variables tuvieran todos los valores, pero cuando algún caso en cualquiera de las 4 variables anteriores tuviera un valor perdido, ya no sería correcto como vemos en el siguiente ejemplo:

HCO31	HCO32	HCO33	HCO34	MediaBicar
22,6	23,8	25,8	25,2	24,35
.	20,9	26,4	24,9	.
23,0	23,8	22,8	23,8	23,35
22,0	24,5	23,9	25,9	24,08
23,8	25,4	25,3	23,7	24,55

Vemos que el segundo caso no ha sido calculado al faltarle el valor en la variable HCO31. Para hacerlo correctamente SPSS dispone de la función MEAN que aparece como una de las opciones del COMPUTE. El procedimiento correcto sería el siguiente:

COMPUTE MediaBicar=MEAN(HCO31,HCO32,HCO33,HCO34).
EXECUTE.

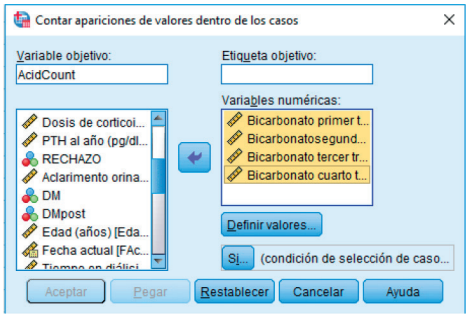
Tras ejecutar la sintaxis el resultado es el siguiente:

HCO31	HCO32	HCO33	HCO34	MediaBicar
22,6	23,8	25,8	25,2	24,35
.	20,9	26,4	24,9	24,07
23,0	23,8	22,8	23,8	23,35
22,0	24,5	23,9	25,9	24,08
23,8	25,4	25,3	23,7	24,55

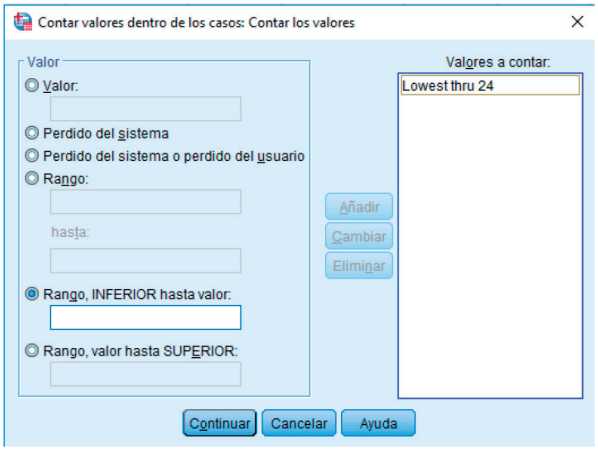
Ahora podemos ver que sí se ha calculado la media del segundo valor.

Otra opción interesante de SPSS es la opción COUNT. Básicamente consiste en indicarle al programa que nos cuente cuántas veces aparece un valor en un conjunto de variables. Por ejemplo, podemos pedirle que nos cuente cuantas veces aparece un valor < 24 en el conjunto de las 4 variables de bicarbonato. Como podemos ver es un paso similar al que hemos hecho previo al SUM para crear la variable "Acidosis" si las ¾ partes de los valores de bicarbonato son inferiores a 24. El procedimiento es el siguiente:

Transformar → Contar valores dentro de los casos. Se nos abrirá el siguiente cuadro de diálogo:



Damos nombre a la nueva variable (sin etiqueta, pues la definiremos mejor tras haberla creado a través de "definir variables"), y seleccionamos las 4 variables a contar. Hacemos Clic sobre el botón "Definir valores" y se nos abrirá otro cuadro de diálogo:



En este cuadro tenemos varias opciones: definir un único valor en la opción "Valor"; que nos cuente cuantos valores están perdidos; contar cuántas veces aparece un valor dentro de un rango (por ejemplo contar cuántas veces aparece un valor entre 22 y 24 de bicarbonato); o que nos cuente cuántas veces aparece un valor inferior o superior a uno que definamos. En este caso queremos que nos cuente cuántas veces aparece un valor inferior a 24 en cada caso en las 4 variables de bicarbonato. Hacemos Clic en el botón "Añadir" y posteriormente en "Continuar". Tras ejecutar la sintaxis el resultado es el siguiente:

HCO31	HCO32	HCO33	HCO34	AcidCount
22,6	23,8	25,8	25,2	2,00
.	.	.	.	,00
23,0	23,8	22,8	23,8	4,00
22,0	24,5	23,9	25,9	2,00
23,8	25,4	25,3	23,7	2,00
20,2	18,6	20,9	21,4	4,00
21,1	27,9	29,5	32,6	1,00
24,8	30,6	30,1	27,9	,00

Podemos ver que en el primer caso existe un valor inferior a 24 en 2 ocasiones (22.6 y 23.8). En el último caso mostrado no hay ningún valor inferior a 24. Pero tiene un problema, que reside en el segundo caso: todos los valores de las 4 variables están perdidos y sin embargo el procedimiento nos da un valor de 0. Realmente es 0 puesto que si contamos cuantos son inferiores a 24 no hay ninguno, es decir, 0, pero no es un valor lógico. En este

caso el procedimiento que hicimos al principio con 4 COMPUTE y después SUM es el correcto puesto que para este caso en concreto el valor debe ser perdido como vemos en el siguiente resultado:

HCO31	HCO32	HCO33	HCO34	SUMA
22,6	23,8	25,8	25,2	2,00
.	.	.	.	.
23,0	23,8	22,8	23,8	4,00
22,0	24,5	23,9	25,9	2,00
23,8	25,4	25,3	23,7	2,00
20,2	18,6	20,9	21,4	4,00
21,1	27,9	29,5	32,6	1,00
24,8	30,6	30,1	27,9	,00

Cuando no hay valores perdidos en ninguna de las variables, ambos procedimientos dan el mismo resultado, siendo más rápido y sencillo el COUNT.

Más difícil todavía: Procedimiento IF

El procedimiento "IF" permite la creación de nuevas variables (o modificación de las ya existentes) mediante expresiones lógicas. La estructura del procedimiento sería la siguiente:

IF (Variable a modificar, cálculo, etc.) Variable nueva (o ya existente) = valor nuevo (variable existente, etc.).

Lo anterior se traduciría como:

Si (variable… =….) entonces (variable…) = ….

Podemos utilizar constantes, variables previas, operadores matemáticos, etc. Dentro de las relaciones podemos usar EQ (igual =), NE (diferente < > ó ~=), LT (menor que <), LE (menor o igual que <=), GT (mayor que >) y GE (mayor o igual que >=). Los términos de conexión son OR (|), AND (&) y NOT (~).

Ejemplo: imaginemos que queremos darle el valor 0 a los pacientes con sexo femenino y presencia de DM, el procedimiento IF sería (el valor de la nueva variable vamos a llamarla Var1):

IF (Sexo=1 AND DM=1) Var1=0.

EXECUTE.

El procedimiento IF no aparece disponible como tal en SPSS y por lo tanto la única forma de ejecutarlo es escribiéndolo directamente a mano en la hoja de sintaxis. Como condiciones, primero debemos tener previamente las variables perfectamente definidas, y segundo, el

procedimiento debe ser exhaustivo, es decir, todas las posibles combinaciones deben estar presentes para "no dejar valores sin opción, sueltos". En el procedimiento del ejemplo anterior no quedan incluidos los pacientes que fuesen varones, los que no fuesen diabéticos ni los valores perdidos al recoger los resultados.

El resultado del operador lógico puede ser un nuevo valor o puede ser igual al valor de una variable ya existente. Por ejemplo, imaginemos que si un paciente tiene sexo femenino y es diabético le damos el valor 0 (ejemplo anterior), pero si es sexo femenino y NO es diabético le damos el valor de la variable DMpost. Sería así:

IF (Sexo=1 AND DM=1) Var1=0.
IF (Sexo=1 AND DM=0) Var1=DMpost.
EXECUTE.

Vamos a poner en práctica el procedimiento IF con nuestro estudio. Vamos a calcular la estimación del filtrado glomerular mediante la fórmula CKD-EPI. La fórmula es la siguiente:

CKD EPI Equation for Estimating GFR on the Natural Scale Expressed for Specified Race, Sex and Standardized Serum Creatinine (From Ann Intern Med 2009;150:604-612, used with permission)

Race	Sex	Serum Creatinine (mg/dL)	Equation
Black	Female	≤0.7	$GFR = 166 \times (Scr/0.7)^{-0.329} \times (0.993)^{Age}$
Black	Female	>0.7	$GFR = 166 \times (Scr/0.7)^{-1.209} \times (0.993)^{Age}$
Black	Male	≤ 0.9	$GFR = 163 \times (Scr/0.9)^{-0.411} \times (0.993)^{Age}$
Black	Male	>0.9	$GFR = 163 \times (Scr/0.9)^{-1.209} \times (0.993)^{Age}$
White or other	Female	≤0.7	$GFR = 144 \times (Scr/0.7)^{-0.329} \times (0.993)^{Age}$
White or other	Female	>0.7	$GFR = 144 \times (Scr/0.7)^{-1.209} \times (0.993)^{Age}$
White or other	Male	≤ 0.9	$GFR = 141 \times (Scr/0.9)^{-0.411} \times (0.993)^{Age}$
White or other	Male	>0.9	$GFR = 141 \times (Scr/0.9)^{-1.209} \times (0.993)^{Age}$

CKD-EPI equation expressed as a single equation: $GFR = 141 \times \min(Scr/\kappa, 1)^{\alpha} \times \max(Scr/\kappa, 1)^{-1.209} \times 0.993^{Age} \times 1.018$ [if female] x 1.159 [if black] where Scr is standardized serum creatinine in mg/dL, κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1.

Supongamos (como así es), que todos los pacientes de nuestro estudio son de raza blanca. Vemos que la fórmula es diferente dependiendo del sexo y de los valores de Creatinina. Vamos a utilizar el valor de la creatinina al año para su cálculo (Cr4). El procedimiento sería el siguiente:

*IF (Cr4<=0.7 AND Sexo=1) CKD_EPI= 144 * (Cr4 / 0.7) ** -0.329 * 0.993**Edad.*
*IF (Cr4>0.7 AND Sexo=1) CKD_EPI4= 144 * (Cr4 / 0.7) ** -1.209 * 0.993**Edad.*
*IF (Cr4 <= 0.9 AND Sexo=0) CKD_EPI4= 141 * (Cr4 / 0.9) ** -0.411 * 0.993**Edad.*
*IF (Cr4 > 0.9 AND Sexo=0) CKD_EPI4= 141 * (Cr4 / 0.9) ** -1.209 * 0.993**Edad.*
EXECUTE.

Para comprobar que lo anterior se ha realizado de manera correcta, realizamos un nuevo LIST VAR con las variables Sexo, Cr4 y CKD_EPI y vemos si los valores son correctos.

En este caso no hay problemas porque todas las variables tienen todos los valores. Si alguno de los valores fuera missing, el procedimiento con ese caso no funcionaría.

Si a los valores missing quisiéramos darle un valor en concreto, debemos utilizar el término $sysmis. Imaginemos que queremos que el programa no tenga en cuenta a los pacientes con sexo femenino y que sean diabéticos mientras que al resto les queremos dar el valor de la variable DMpost; el procedimiento sería el siguiente:

```
IF (Sexo=1 AND DM=1) Var1=$sysmis.
EXECUTE.
```

Lógicamente el procedimiento anterior no es exhaustivo puesto que no tiene en cuenta a los pacientes de sexo=0, a los pacientes con DM=0 ni a los perdidos en una y otra variable (si los hubiera). El procedimiento completo sería:

```
IF (Sexo=1 AND DM=1) Var1=$sysmis.
IF (Sexo=0 OR DM=0) Var1=DMpost.
IF (MISSING (Sexo) OR MISSING (DM)) Var1=DMpost.
EXECUTE.
```

Imaginemos que queremos crear la variable "DMO" donde un paciente presenta osteoporosis (definido con el valor 2) cuando tiene osteoporosis en columna o en cadera (es decir, en cualquiera de las 2), o tiene densitometría normal (definido con el valor 0) cuando es normal tanto en columna como en cadera (en las 2 a la vez). En el resto de circunstancias lo consideraremos como osteopenia (definido con el valor 1). ¿Cómo crearíamos esta nueva variable? Debemos tener en cuenta que si en alguna de las 2 variables (columna o cadera) hay un valor perdido, el valor de la nueva variable varía, de tal forma que si es perdido en columna y en cadera presenta osteoporosis, el resultado será osteoporosis (hemos dicho que presenta osteoporosis si ésta aparece en cualquiera de las 2). Pero si es perdido en columna y en cadera es inferior a 2, el resultado debe ser perdido puesto que no sabemos realmente qué valor hubiéramos podido encontrar en el perdido. La sintaxis sería la siguiente:

```
IF (DMOColumna=2 OR DMOCadera=2) DMO=2.
IF (DMOColumna=0 AND DMOCadera=0) DMO=0.
IF (DMOColumna=1 AND DMOCadera<2) DMO=1.
IF (DMOColumna<2 AND DMOCadera=1) DMO=1.
IF (MISSING(DMOColumna) AND MISSING(DMOCadera)) DMO=$SYSMIS.
IF (MISSING(DMOColumna) AND DMOCadera<2) DMO=$SYSMIS.
IF (DMOColumna<2 AND MISSING(DMOCadera)) DMO=$SYSMIS.
```

Con el primer IF definimos la presencia de osteoporosis, mientras que con el segundo definimos la presencia de densitometría normal. Con el resto definimos la presencia de osteopenia, indicando que el resultado sea perdido ($SYSMIS) cuando no podamos asegurar si un caso puede tener osteoporosis al tener un valor perdido en alguna de las variables (o en las 2) y en la otra el resultado encontrado sea inferior a 2.

RECODIFICACIÓN DE VARIABLES (PROCEDIMIENTO RECODE)

Con el procedimiento RECODE podemos recodificar una variable en otra variable nueva o bien en la misma. Se utiliza para "agrupar" valores o para darle valores distintos a los ya existentes. Por ejemplo, en nuestro estudio queremos crear una nueva variable que nos agrupe a los pacientes en los distintos estadios de función renal. Podríamos crear 2 variables, una que divida la muestra en pacientes con mayor o menor de 60 ml/min de filtrado glomerular (60 ml/min es el punto de corte que se utiliza para considerar que un paciente tiene insuficiencia renal crónica). La sintaxis es fácil, ya lo hemos visto:

COMPUTE IRC_CKD=CKD_EPI < 60.

EXECUTE.

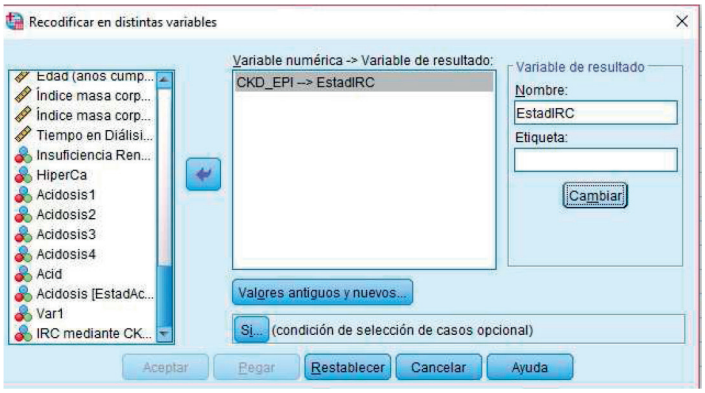
Pero en este caso lo que nos interesa es crear una nueva variable que divida los pacientes en los distintos estadios de IRC de esta manera: Mayor de 60 sin IRC; entre 30-60 Estadio 3; entre 15-30 Estadio 4; y <15 Estadio 5.

Debemos tener en cuenta que SPSS "lee" de izquierda a derecha, de manera que el valor que aparezca en un intervalo no será leído en el siguiente. Por ejemplo si ponemos (15-30) y (30-60) …, el valor 30 entrará a formar parte del primer intervalo.

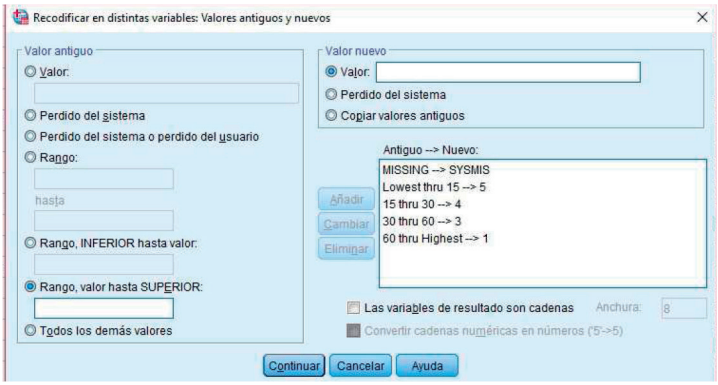
Vamos a ver cómo se hace.

En primer lugar vamos a la opción:

Transformar → Recodificar en distintas variables…, y se nos abrirá el siguiente cuadro de diálogo:



Seleccionamos la variable a recodificar (CKD_EPI) y en la casilla *Nombre* escribimos el valor de la nueva variable y hacemos Clic en *Cambiar*. A continuación hacemos Clic sobre el botón *Valores antiguos y nuevos*… y se nos abrirá otro cuadro de diálogo donde iremos dando el valor a los distintos estadios. En primer lugar seleccionamos los *Valores perdidos por el sistema o por el usuario* y los consideramos como *Valores perdidos del Sistema* para que SPSS no los tenga en cuenta.



En la parte izquierda vamos seleccionando los valores que servirán como punto de corte, le otorgamos un *Valor nuevo* y hacemos Clic sobre *Añadir*, posteriormente a *Continuar* y luego a pegar. Antes de pegar debemos seleccionar la opción *Todos los demás valores* y darles el valor *Perdido del sistema* para que todos los demás valores que no cumplan estos criterios que hemos especificados no nos los meta en ningún intervalo. El resultado de la sintaxis será la siguiente:

RECODE CKD_EPI (MISSING=SYSMIS) (Lowest thru 15=5) (15 thru 30=4) (30 thru 60=3) (60 thru Highest=1) (ELSE=SYSMIS) INTO EstadIRC.

EXECUTE.

Por último tenemos que Definir las propiedades de la nueva variable. Vemos que no hay ningún paciente con Estadio 5; pero debemos crearlo porque si volvemos a utilizar de nuevo la base de datos con pacientes nuevos y hubiese alguno con Estadio 5 no aparecerá si no lo hemos Definido previamente.

Con este procedimiento lo que se consigue es dividir la muestra en intervalos, es decir, transforma una variable cuantitativa continua en una variable categórica con varias categorías.

Hay que recordar que este procedimiento no es el adecuado para crear una variable binaria, sino el COMPUTE que hemos visto anteriormente, porque con el RECODE corremos el riesgo de que algunos valores que son Missing (como los User Missing) los tenga en cuenta y les otorgue valor.

Con el procedimiento RECODE también se pueden transformar variables en formato "letra" a formato numérico. Para ello se necesita que la letra vaya entre '_'. Por ejemplo, si el sexo lo hemos codificado como M para masculino y F para femenino y le queremos dar el valor 0 y 1, el procedimiento sería:

RECODE Sexo (MISSING=SYSMIS) (' M'=0) ('F'=1)

(ELSE=SYSMIS) INTO Sex.

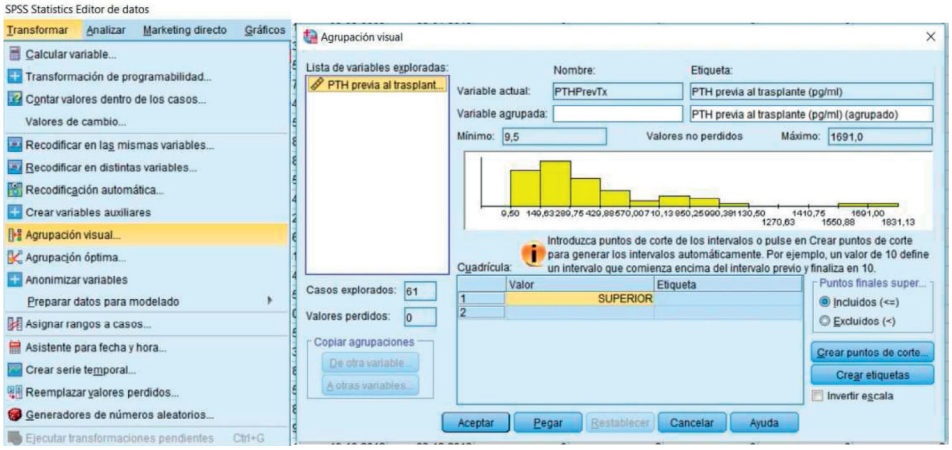
EXECUTE.

RECODIFICACIÓN VISUAL

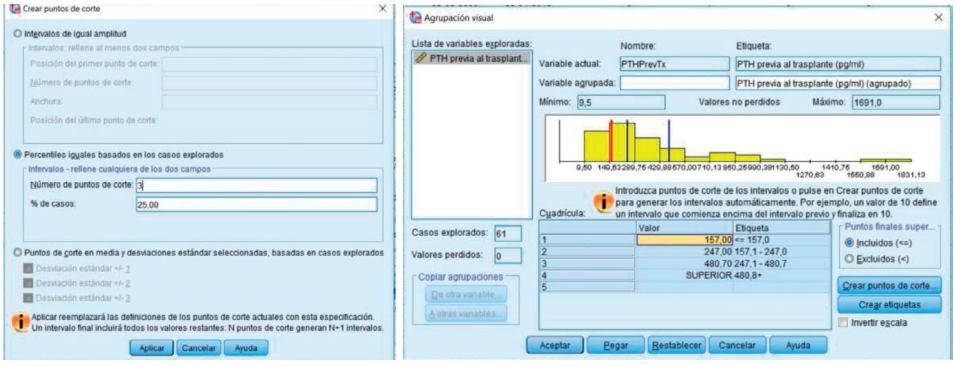
Con la recodificación visual podemos transformar una variable cuantitativa en una variable categórica con varias categorías a través de puntos de corte a "mano alzada" o en intervalos de igual tamaño, por ejemplo con percentiles, cuartiles, etc.

Para ello seguimos el siguiente paso:

Transformar → Agrupación visual. Se nos abrirá el cuadro de diálogo de la derecha:



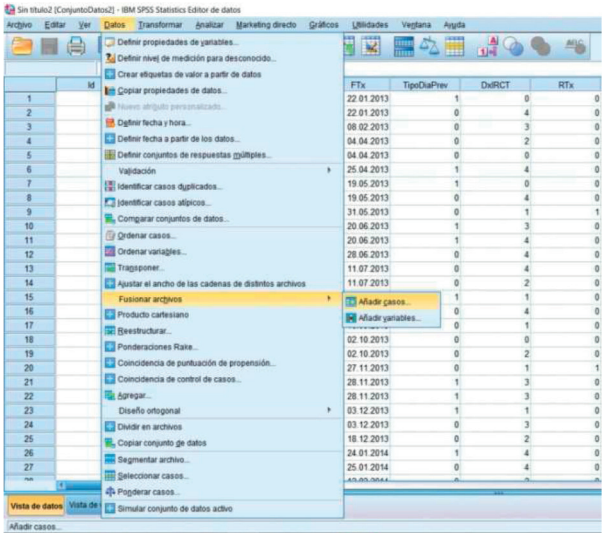
Hacemos Clic sobre el botón "Crear puntos de corte" y se abrirá el siguiente cuadro de diálogo. En él podemos crear intervalos de igual amplitud o crear puntos de corte a través por ejemplo de los percentiles; si ponemos 3 puntos de corte se crearán 4 grupos divididos por cuartiles. Tras darle al botón "Aplicar", volvemos al cuadro anterior donde se habrán creado los puntos de corte en líneas de colores. Si situamos el cursor sobre esas líneas, podemos moverlas a nuestro antojo. También tenemos la opción de crear las etiquetas directamente.



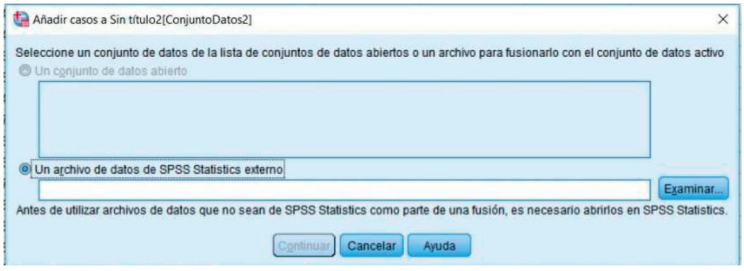
PROCEDIMIENTO "ADD FILES"

Con este procedimiento vamos a añadir filas de otra base de datos, (cuyos nombres de las variables deben ser iguales), sobre la base de datos actual. Es útil para estudios multicéntricos donde cada centro va a tener una copia exacta de la original que tendremos que unificar en una sola tabla de datos. El procedimiento es el siguiente:

Datos → Fusionar archivos → Añadir casos…



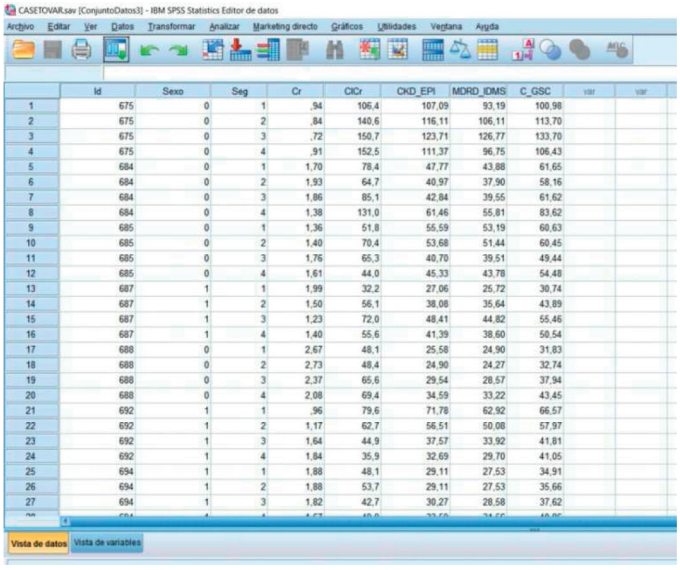
Se abrirá un cuadro de diálogo donde seleccionaremos la ubicación de la base de datos a importar a través del botón "Examinar".



PROCEDIMIENTO "CASETOVAR"

Este procedimiento se utiliza para transformar casos en variables. Cuando realizamos estudios de seguimiento de pacientes, por ejemplo en una consulta médica, las bases de datos suelen registrar cada visita como si se tratase de un nuevo paciente cuando en realidad es el mismo que ha venido varias veces. Si queremos comparar distintas variables de seguimiento de los pacientes debemos transformar estos casos en variables. Para este ejemplo vamos a utilizar la base de datos "CASETOVAR.sav" de la carpeta "Bases de datos".

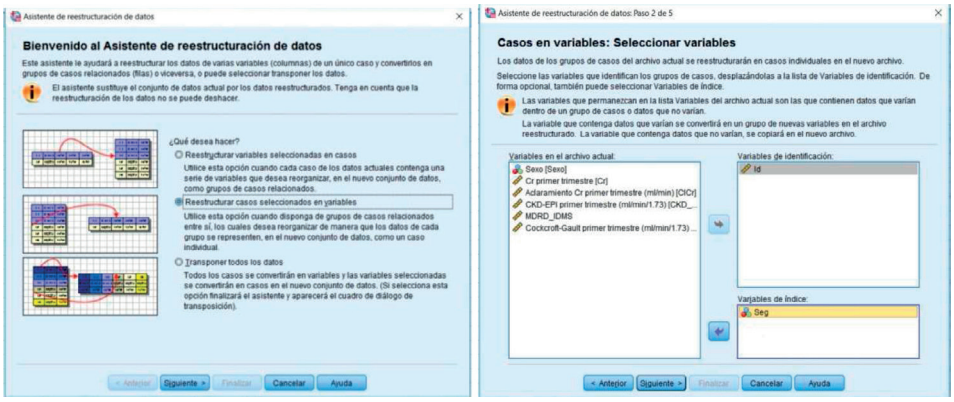
En la siguiente imagen donde cada paciente viene identificado con un número en la variable "Id" podemos ver que el paciente número 675 viene registrado 4 veces que corresponde a cada una de sus visitas a la consulta. El número de la visita aparece en la variable "Seg" con un número del 1 al 4 dependiendo de si es la primera, segunda, tercera o cuarta visita. Si por ejemplo queremos analizar si varía el valor de "Cr" en las sucesivas visitas a la consulta en cada paciente, debemos transformar los casos en variables de tal forma que sólo exista un caso por registro. Para ello debemos tener perfectamente identificado cada sujeto (en este caso con un número en la variable "Id") y una variable que nos indique el número de seguimiento, en este caso la variable "Seg".



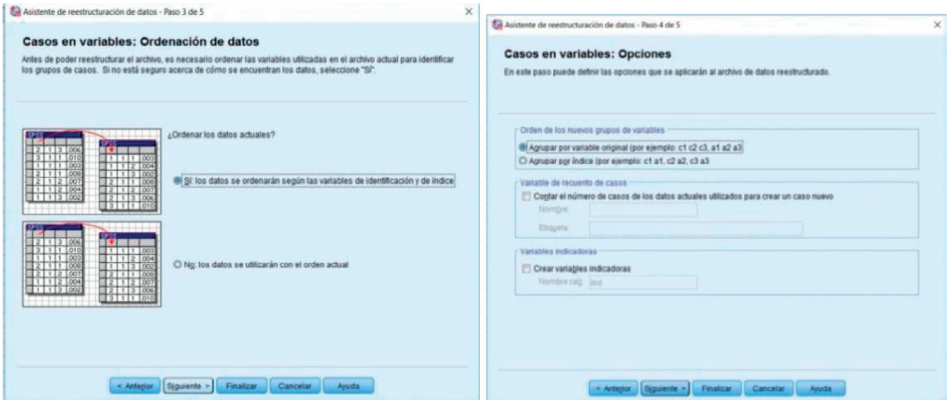
	Id	Sexo	Seg	Cr	ClCr	CKD_EPI	MDRD_IDMS	C_GSC
1	675	0	1	,94	106,4	107,09	93,19	100,98
2	675	0	2	,84	140,6	116,11	106,11	113,70
3	675	0	3	,72	150,7	123,71	126,77	133,70
4	675	0	4	,91	152,5	111,37	96,75	106,43
5	684	0	1	1,70	78,4	47,77	43,88	61,65
6	684	0	2	1,93	64,7	40,97	37,90	58,16
7	684	0	3	1,86	85,1	42,84	39,55	61,62
8	684	0	4	1,38	131,0	61,46	55,81	83,62
9	685	0	1	1,36	51,8	55,59	53,19	60,63
10	685	0	2	1,40	70,4	53,68	51,44	60,45
11	685	0	3	1,76	65,3	40,70	39,51	49,44
12	685	0	4	1,61	44,0	45,33	43,78	54,48
13	687	1	1	1,99	32,2	27,06	25,72	30,74
14	687	1	2	1,50	56,1	38,08	35,64	43,89
15	687	1	3	1,23	72,0	48,41	44,82	55,46
16	687	1	4	1,40	55,6	41,39	38,60	50,54
17	688	0	1	2,67	48,1	25,58	24,90	31,83
18	688	0	2	2,73	48,4	24,90	24,27	32,74
19	688	0	3	2,37	65,6	29,54	28,57	37,94
20	688	0	4	2,08	69,4	34,59	33,22	43,45
21	692	1	1	,96	79,6	71,78	62,92	66,57
22	692	1	2	1,17	62,7	56,51	50,08	57,97
23	692	1	3	1,64	44,9	37,57	33,92	41,81
24	692	1	4	1,84	35,9	32,69	29,70	41,05
25	694	1	1	1,88	48,1	29,11	27,53	34,91
26	694	1	2	1,88	53,7	29,11	27,53	35,66
27	694	1	3	1,82	42,7	30,27	28,58	37,62

El procedimiento es el siguiente:

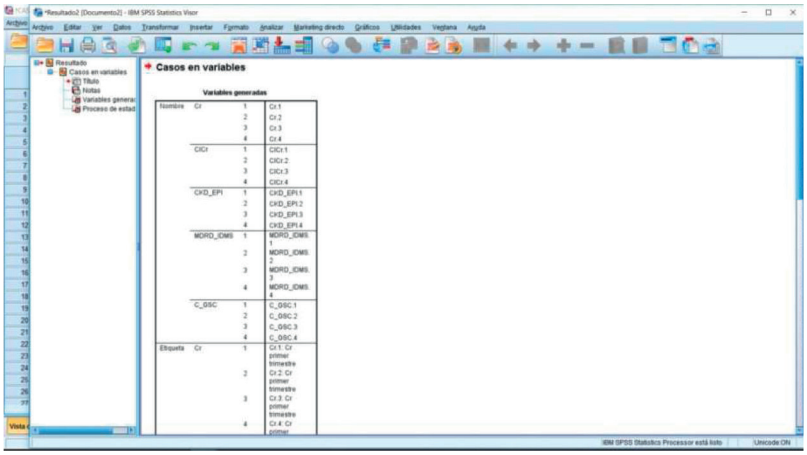
Datos → Reestructurar; se nos abrirá el siguiente cuadro de diálogo donde seleccionaremos la opción "Reestructurar casos seleccionados en variables"



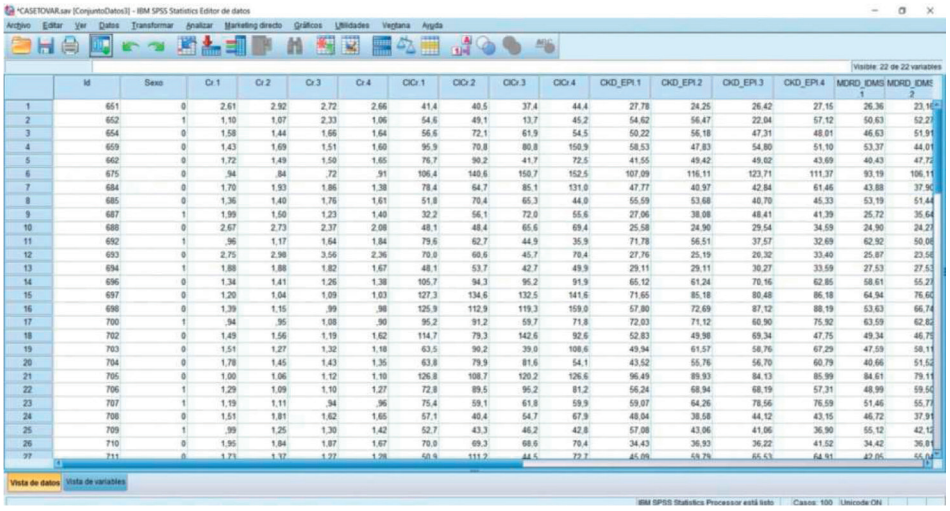
En el cuadro de diálogo de la derecha, en el recuadro "Variables de identificación" seleccionaremos la variable que nos sirve para identificar cada sujeto, en este caso la variable "Id". En el recuadro "Variables de índice" seleccionaremos la variable que nos indica el seguimiento de los pacientes, en este caso la variable "Seg". Tras hacer Clic en "Siguiente" se nos abren los siguientes cuadros de diálogos:



En el cuadro de la izquierda le indicamos que "queremos ordenar los datos según las variables de identificación y de índice". Al hacer Clic en "Siguiente" se abre el cuadro de la derecha, donde seleccionamos la opción "Agrupar por variable original". Tras finalizar, en la hoja de resultado nos aparecerá una tabla resumen de la manera en como se ha reestructurado la hoja de datos.

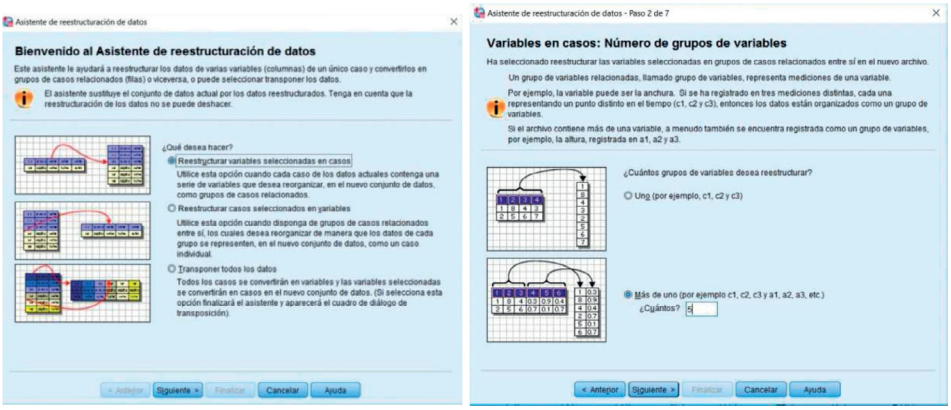


La hoja de datos resultante final es la siguiente donde podemos observar que sólo hay un caso por paciente, y las variables de seguimiento aparecen ordenas. Ahora ya podemos comparar las distintas variables de cada paciente en cada seguimiento.



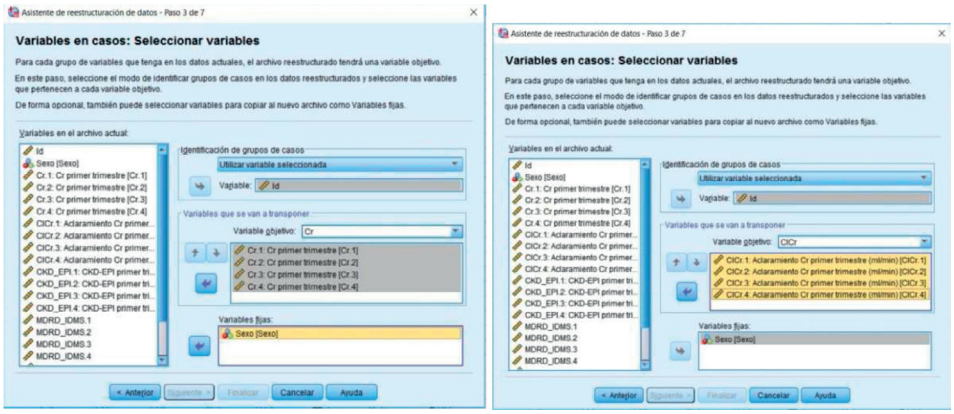
PROCEDIMIENTO "VARTOCASE"

Este procedimiento es el inverso al anterior, donde transformamos variables en casos. En el primer cuadro de diálogo seleccionamos en este caso la primera opción "Reestructurar variables seleccionadas en casos". En el paso número 2 (cuadro siguiente de la derecha), seleccionamos la opción "Más de uno (por ejemplo c1, c2, c3 y a1, a2, a3, etc.)", escribiendo el número de grupos a crear. En el ejemplo que estamos utilizando debemos crear agrupar las variables "Cr", "ClCr", "CKD-EPI", "MDRD" y "C-G_SG", por tanto indicamos en "¿Cuántos?" el valor 5.

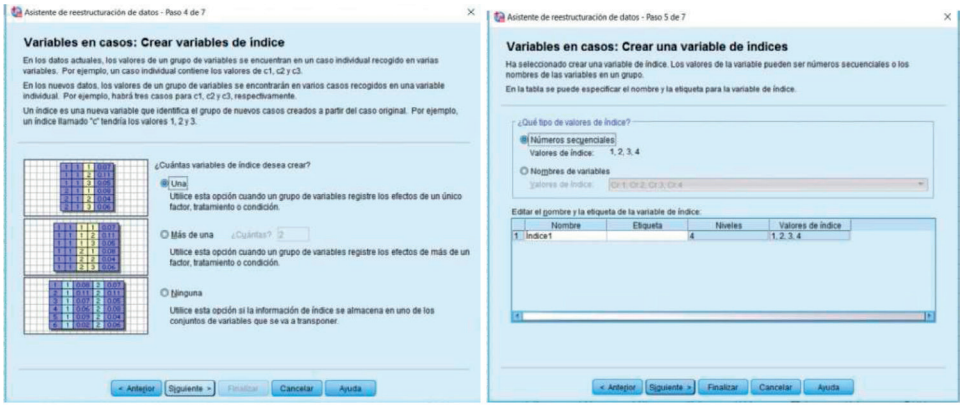


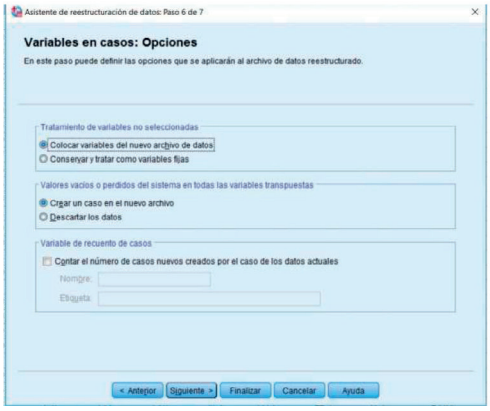
En el siguiente paso seleccionamos la variable que nos va a servir para identificar cada sujeto, en este caso la variable "Id". Si el valor de alguna variable es el mismo independientemente del momento del seguimiento lo seleccionamos en el recuadro "Variables fijas";

por ejemplo, el Sexo no varia con el seguimiento pues es siempre el mismo en cada sujeto, por tanto lo desplazamos hacia este recuadro. En el apartado "Variables que se van a transponer" en la opción "Variable objetivo" vamos escribiendo el nombre de la nueva variable de agrupación (el nombre por defecto es "Trans1", "Trans2, …) y arrastramos hacia el recuadro del medio las variables correspondientes a esa agrupación. Por ejemplo, la variable "Cr" va a agrupar las variables "Cr1", "Cr2", "Cr3" y "Cr4". Nombramos y agrupamos todas las variables.



Tras lo anterior indicamos que nos cree una sola variable de índice y hacemos Clic en "Siguiente" en todos los demás pasos. El resultado será una tabla igual a la que teníamos antes del procedimiento "CASETOVAR".

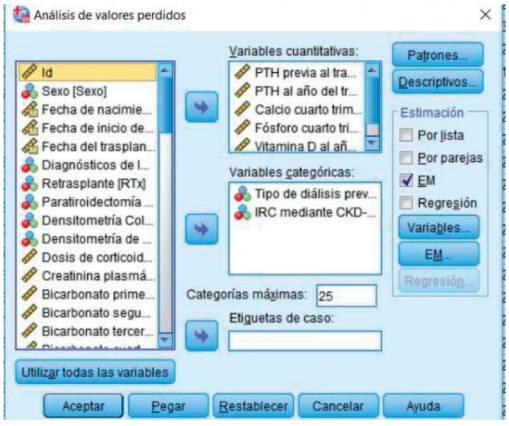




ESTIMACIÓN DE VALORES PERDIDOS

Cuando realizamos análisis estadísticos, es deseable que el número de valores perdidos en las variables sea el menor posible. Un paso previo antes de realizar cualquier análisis es evaluar cuántos valores perdidos tengo en mi base de datos. Además, hay que saber que cuando SPSS realiza análisis estadísticos con más de una variable, si algún caso presenta un valor perdido en alguna de las variables a analizar, directamente ese caso no lo tiene en cuenta y no entra en el análisis. Cuando hacemos análisis estadísticos avanzados donde entran varias variables (análisis multivariantes que veremos más adelante en los capítulos de regresión), el tamaño de la muestra para estimar los resultados se puede ver reducido si existen valores perdidos en las distintas variables porque va a desechar todos los casos en los que haya un valor perdido. En estas circunstancias, nos podría interesar "estimar" el valor que tendría un caso en una determinada variable. SPSS tiene implementado el procedimiento de análisis de valores perdidos, donde no sólo vamos a poder ver cuantos valores tenemos perdidos, sino también estimar esos valores perdidos. El procedimiento es el siguiente:

Analizar → Análisis de valores perdidos. Se nos abrirá el siguiente cuadro de diálogo donde vamos a seleccionar las variables a analizar, generalmente todas. Posteriormente seleccionamos la opción "EM".



La sintaxis es la siguiente, la cual debemos modificar y escribir al final antes del paréntesis final, la dirección en la que queremos que se nos guarde la nueva tabla de datos con todos los valores estimados:

```
MVA VARIABLES=PTHPrevTx PTH4 Ca4 P4 VitD12 TiempHD Sexo TipoDiaPrev
/MAXCAT=25
/CATEGORICAL=Sexo TipoDiaPrev
/EM(TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25
OUTFILE='C:\Users\Admin\Desktop\Curso\TREM.sav').
```

Truco: si no sabemos cómo escribir la dirección en la que queremos que se guarde la nueva tabla, podemos hacer lo siguiente: en la tabla de datos abierta, hacemos el siguiente procedimiento:

Archivo → Guardar como → Pegar; seleccionamos la ubicación que queramos, generalmente el mismo sitio en el que estemos trabajando (disco duro, pendrive, etc.); nos dirá si queremos cambiar la tabla y le decimos que sí (sólo se va a pegar en la sintaxis, hasta que no la ejecutemos no pasa nada). Una vez pegada en la sintaxis, cambiamos el nombre de la tabla por el que queramos (en el ejemplo se ha llamado "TREM") y lo pegamos justo después del 25 y antes del paréntesis.

Tras ejecutar la sintaxis obtenemos lo siguiente:

Estadísticos univariados

	N	Media	Desviación estándar	Perdidos		Número de extremos[a]	
				Recuento	Porcentaje	Menor	Mayor
PTHPrevTx	61	386,603	347,3255	0	,0	0	3
PTH4	55	147,109	112,9385	6	9,8	0	4
Ca4	61	9,739	,5466	0	,0	1	1
P4	61	3,072	,6435	0	,0	0	2
VitD12	38	16,258	6,8176	23	37,7	0	0
TiempHD	61	72,482	68,5711	0	,0	0	5
Sexo	61			0	,0		
TipoDiaPrev	61			0	,0		

a. Número de casos fuera del rango (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

En esta primera tabla aparece un Descriptivo de las variables en cuestión con la Media, etc., y el Recuento y Porcentaje de los valores perdidos de cada una de ellas, además de los valores extremos. En las siguientes tablas, nos muestra el valor de la media y desviación estándar estimada una vez sustituidos los valores perdidos por los nuevos valores estimados. En la última tabla que se presenta, aparece debajo un valor de significación de Chi-cuadrado que nos indica si los valores perdidos se han producido por azar o siguen algún patrón de pérdida (por ejemplo, imaginemos que sólo se pierden los valores de los pacientes varones). Si no es estadísticamente significativo, indica que las pérdidas se producen por azar, de manera aleatoria, que es precisamente los que nos interesa.

Resumen de medias estimadas

	PTHPrevTx	PTH4	Ca4	P4	VitD12	TiempHD
Todos los valores	386,603	147,109	9,739	3,072	16,258	72,482
EM	386,603	148,007	9,739	3,072	15,918	72,482

Resumen de desviaciones estándar estimadas

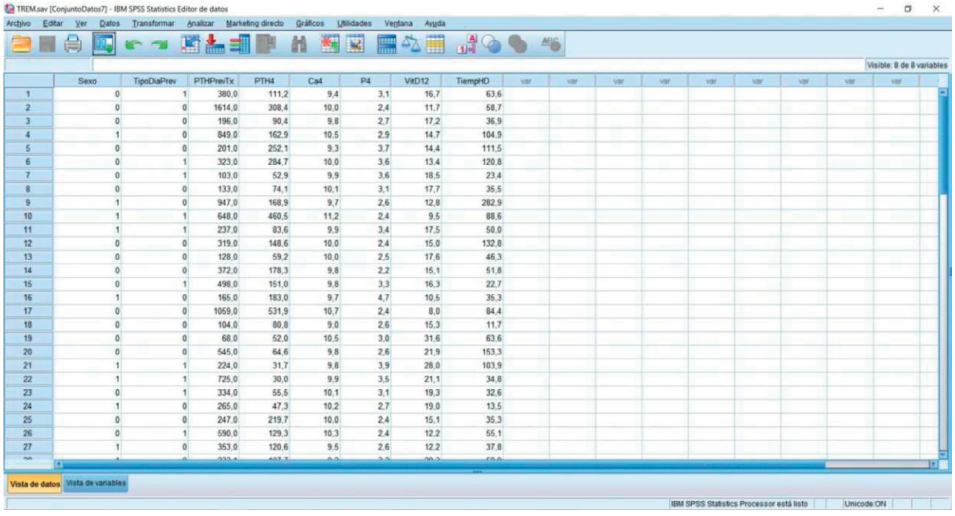
	PTHPrevTx	PTH4	Ca4	P4	VitD12	TiempHD
Todos los valores	347,3255	112,9385	,5466	,6435	6,8176	68,5711
EM	347,3255	114,1600	,5466	,6435	6,9288	68,5711

Medias marginales estimadas[a]

PTHPrevTx	PTH4	Ca4	P4	VitD12	TiempHD
386,603	148,007	9,739	3,072	15,918	72,482

a. Prueba MCAR de Little: Chi-cuadrado = 11,680, DF = 14, Sig. = ,632

Si abrimos la base de datos nueva creada, aparecerán los valores estimados en los casos que tuvieran valores perdidos previamente.

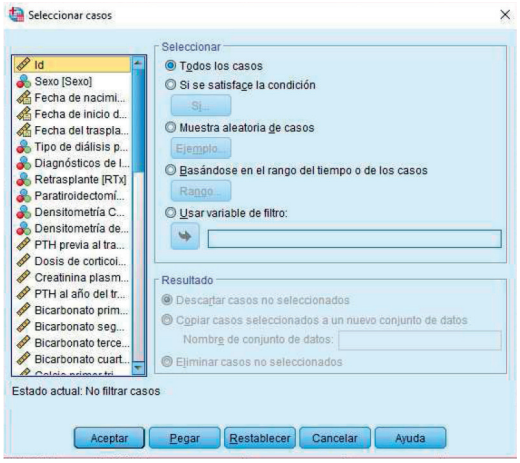


PROCEDIMIENTO SELECCIÓN DE CASOS

Con este procedimiento lo que conseguimos es seleccionar sólo una parte de la muestra en función de unos criterios que nosotros establezcamos. Estos criterios pueden ser una determinada variable, el resultado de una ecuación, una expresión lógica, etc.

Por ejemplo, imaginemos que queremos analizar los datos sólo de los pacientes varones de nuestro estudio. El procedimiento sería el siguiente: en la lista de opciones seleccionar la opción:

Datos → Seleccionar casos… Se nos abrirá una página de diálogo como la siguiente:



Vemos que tenemos las opciones: *Si se satisface la condición, Muestra aleatoria de casos, Basándose en el rango del tiempo o de los casos, y Usar variables de filtro. La opción que vamos a usar prácticamente siempre es Si se satisface la condición.* Si seleccionamos esta opción y posteriormente hacemos Clic en *Si…*, se nos abrirá el siguiente cuadro de diálogo:

Vemos que es un cuadro de diálogo igual al del COMPUTE. En la parte de arriba vemos que pone "Seleccionar casos: Si la opción…". Podemos escribir la opción que deseemos (que una variable tenga un determinado valor, la combinación de varias variables, una ecuación,…). En nuestro caso que la variable Sexo sea igual a 0 (Hombre). La sintaxis sería:

USE ALL. COMPUTE filter_$=(Sexo = 0). VARIABLE LABELS filter_$ 'Sexo = 0 (FILTER)'. VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0). FILTER BY filter_$. EXECUTE.

La ejecutamos tras pegarla y vemos que en la hoja de datos habrá casos que aparezcan tachados. Son los casos con Sexo=0 que no se van a tener en cuenta para los procedimientos que realicemos de aquí en adelante.

¡Cuidado!, si queremos volver a realizar procedimientos estadísticos en el total de la muestra debemos desactivar el filtro con la opción "Seleccionar Todos los Casos".

CREACIÓN Y EDICIÓN DE TABLAS

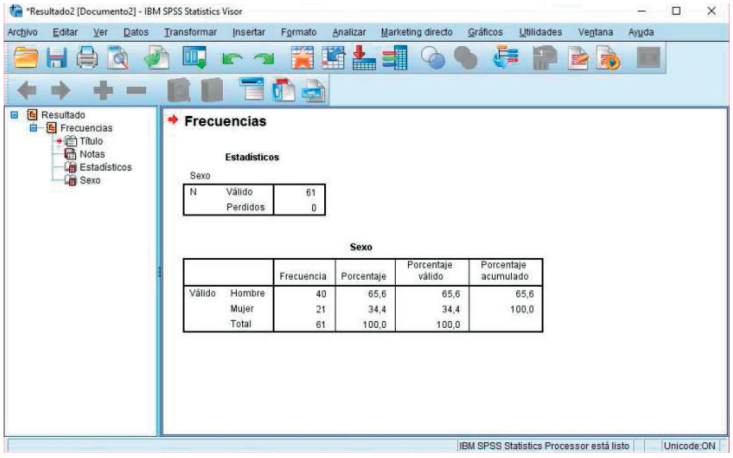
Comienza la parte más personalizada de SPSS: la creación y edición de Tablas (y gráficos que veremos en el siguiente apartado).

En este caso no hay una regla estricta y cada uno puede editar una tabla a su gusto. Hay que decir que las tablas que elabora SPSS son demasiado sencillas y habrá casi siempre que editarla. También hay que decir que en la mayoría de las ocasiones las tablas de SPSS no van a ser las que vamos a utilizar para una publicación y es preferible tomar los datos de cada una de ellas y elaborar una personal en otro programa, como Word.

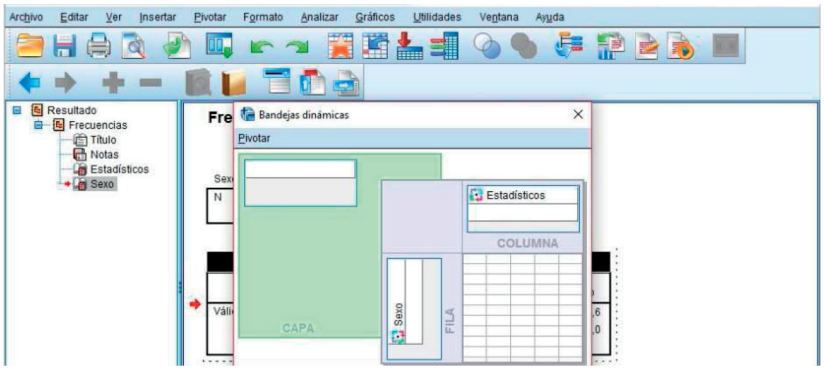
Vamos a crear una tabla cualquiera y ver qué puede realizarse sobre ella. Vamos a crear la tabla de resumen de la variable Sexo mediante el procedimiento:

Analizar → Estadísticos Descriptivos → Frecuencias… y seleccionamos la variable Sexo. Se nos abre el siguiente cuadro de diálogo:

A la derecha tenemos varias opciones: Si hacemos Clic en *Estadísticos…* se nos abre otro cuadro de diálogo donde podemos seleccionar las opciones *Mediana, Media, etc.*, que en este caso no nos interesa porque es una variable categórica. Si hacemos Clic sobre *Gráficos…* se nos abre otro cuadro de diálogo por si queremos que nos haga un gráfico sobre la variable seleccionada. En la opción *Formato* podemos pedirle a SPSS que los resultados los dé en orden ascendente, descendente… Las opciones *Estilo…* y *Simular muestreo…* no nos interesan. Abajo a la izquierda aparece seleccionada por defecto la opción Mostrar tablas de frecuencias. En este caso como es una variable categórica la dejamos seleccionada; si fuese una variable cuantitativa debemos deseleccionarla porque el resultado es una tabla con la frecuencia de cada uno de los resultados. Tras ejecutar la sintaxis se nos abre una hoja de resultados con la tabla siguiente:



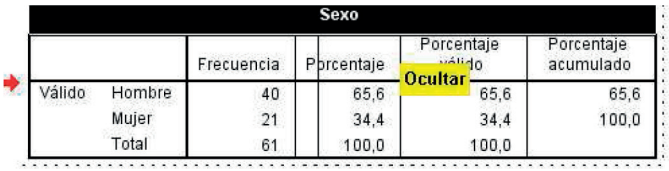
Sobre esta tabla ahora podemos hacer las modificaciones pertinentes. Para ello debemos hacer doble Clic sobre ella para que nos aparezca el editor. La primera vez que lo hagamos se nos abre una opción como la siguiente:



Sobre este cuadro de diálogo podemos intercambiar las variables existentes para que aparezcan en la fila o en la columna. Para ello tan sólo tenemos que hacer Clic sobre la variable y sin soltar el botón del ratón arrastrarla hacia donde queramos. Si este cuadro no

aparece al hacer doble Clic sobre la tabla, podemos hacer que aparezca haciendo Clic en la opción de la cinta de opciones Pivotar → Bandejas dinámicas. O bien haciendo Clic sobre la opción *Transponer filas por columnas.*

En alguna ocasión no nos va a interesar que aparezcan algunas de las opciones; por ejemplo, en los porcentajes sólo nos puede interesar que aparezca el *porcentaje válido* que es el que tiene en cuenta el cálculo del porcentaje con los valores perdidos, y que no aparezca *Porcentaje* ni *Porcentaje Acumulado*. Para ello nos situamos con la flecha sobre el límite del recuadro en cuestión hasta que aparezca una flecha con doble sentido (↔) y arrastramos hacia la izquierda hasta que aparezca en el recuadro que nos informa sobre la longitud del cuadro que estamos creando, la palabra "Ocultar"; tras ello, el recuadro con la información en cuestión ya no aparecerá.



Si no queremos que aparezca el resultado "Total" en nuestra tabla, primero deberemos hacer una Transposición de filas en columna, para que el "Total" aparezca en columna y luego realizar el procedimiento anterior hasta que aparezca "Ocultar". Posteriormente habrá que volver a realizar una Transposición de filas en columna para volver al estado original.

Podemos modificar el nombre de las variables o de las opciones que aparecen en los recuadros, así como el tamaño de letra, tipo de letra, negrita, colorear el recuadro, etc., igual como si fuera una tabla de Word.

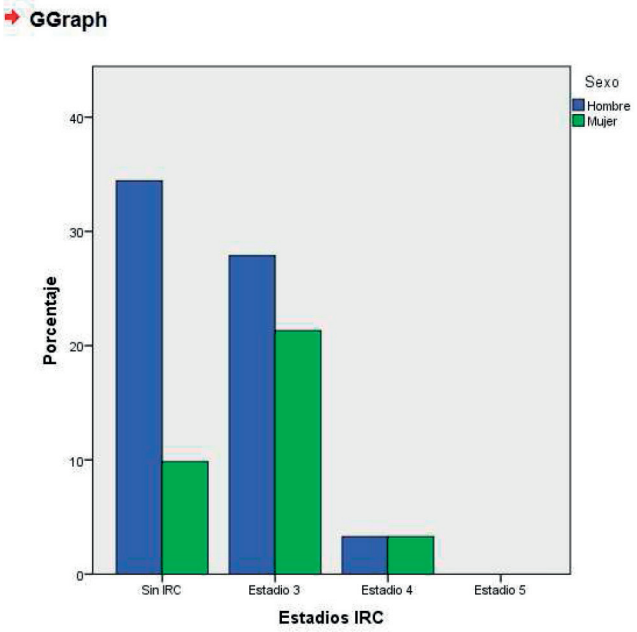
CREACIÓN Y EDICIÓN DE GRÁFICOS

Con los gráficos sucede algo similar que con las Tablas. Es algo personalizado, con multitud de opciones que pueden ser modificables a gusto del consumidor. Vamos a realizar un pequeño ejemplo con el gráfico obtenido al crear un diagrama de barras entre el sexo y los estadios de IRC. Hacemos Clic sobre la opción de la cinta de opciones:

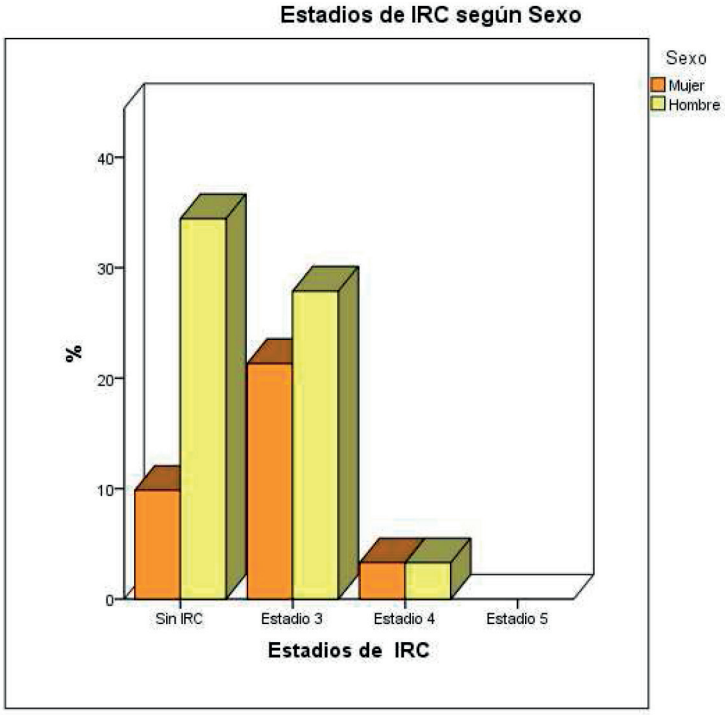
Gráficos → Generado de Gráficos…

La primera vez que lo hacemos nos informa de que las distintas variables deben estar correctamente definidas para que los gráficos puedan realizarse. Es un cuadro de diálogo bastante intuitivo donde podemos seleccionar el tipo de gráficos que queramos realizar. Lo seleccionamos y lo arrastramos hacia el cuadro blanco de arriba. Aparecerá un cuadro como el siguiente:

En este caso seleccionamos el segundo tipo de gráficos dentro de los existentes en los de *Barra*. Arrastramos la variable Sexo hacia el recuadro que hay en la parte superior derecha *Agrupar en X...* La variable Estadios de IRC la arrastramos hacia el eje de abscisas. A la derecha aparece otro cuadro de diálogo que nos permite elegir el tipo de información que debe recoger el gráfico (recuento total, porcentaje válido, etc.). En este caso seleccionamos *Porcentaje* (¿), hacemos Clic en *aplicar* y posteriormente en *pegar*. Ejecutamos toda la sintaxis resultante y el gráfico obtenido es:



Vemos que el gráfico obtenido no es demasiado bonito (colores azul y verde, fondo gris….). Esto ahora podemos editarlo a nuestro gusto haciendo doble Clic sobre el mismo, abriéndose el editor de gráficos. Podemos seleccionar cualquier cosa que queramos: la leyenda, cambiar el color de los rectángulos, el tamaño de los mismo, seleccionar todos los rectángulos de una opción (haciendo Clic una sola vez) o sólo un rectángulo (haciendo Clic dos veces seguidas sobre un rectángulo), añadir un Título, y pie de figura, un cuadro de texto, etc. Editar el gráfico resultante hasta obtener uno lo más parecido posible al siguiente:



DESCRIPCIÓN DE VARIABLES

El primer paso previo al análisis de datos de un estudio de investigación es describir la muestra de la cual proceden los datos; e incluso en algunos tipos de estudios (Transversales) puede ser el único objetivo del mismo, es decir, describir lo que "vemos" (Estudios Descriptivos) para informar sobre una situación concreta o para posteriormente diseñar otro tipo de estudios con el fin de comprobar una sospecha (hipótesis).

Generalmente estamos acostumbrados a ver en los estudios que las variables se describen como la media ± desviación estándar o número de caso y porcentaje. Sin embargo, éstas no son siempre las maneras más correctas de descripción de una variable en concreto, e incluso, puede que no sea ni la manera más correcta de escribirlo.

Previo a lo anterior debemos saber en primer lugar los tipos de variables que existen. Vamos a tener variables categóricas (no se pueden medir) y variables cuantitativas (se pueden medir).

- Variables cuantitativas: son variables que pueden ser medidas o contadas. A su vez pueden ser discretas o continuas. Ejemplos:

  – Discretas: número de embarazos/abortos, número de ingresos hospitalarios,…

  – Continuas: Creatinina, Urea, edad,…

- Variables categóricas: son variables que no pueden ser medidas con ningún instrumento de medida, por ejemplo el sexo. Estas variables a su vez se van a dividir en nominales (no siguen un orden) y en ordinales (siguen un orden). Las nominales además pueden ser binarias o tener más de una categoría. Ejemplos:

  – Nominales:

    Binarias: Sexo (hombre/mujer), HTA (Sí/No), IRC (Sí/No).

    Más de una categoría: Diagnóstico de IRC (Nefropatía Diabética, Poliquistosis, Nefritis, etc).

  – Ordinales: Estadios de IRC (Estadio 1, 2, 3, 4 y 5), Consumo tabaco (Fumador, Exfumador, No fumador).

Tener en cuenta el tipo de variable es necesario para una correcta definición de las mismas en SPSS puesto que el programa no va a poder realizar los mismos procesos estadísticos con unas variables o con otras. Hay algo a tener en cuenta, a SPSS no le gusta trabajar con letras, por tanto es preferible recodificar las variables categóricas a números.

Por ejemplo, no vamos a recoger los datos de sexo como *Hombre* o *Mujer*, sino que es preferible recodificarlo o recogerlo inicialmente con valores numéricos (0, 1) e indicarle a SPSS que se trata de variables categóricas.

DESCRIPCIÓN DE VARIABLES CUANTITATIVAS

Como se comentaba anteriormente, estos datos estamos habituados a verlos en los estudios como media ± desviación estándar. Sin embargo, no siempre es la manera más correcta de hacerlo. E incluso, en algunas publicaciones no se permite incorporar el signo "±" y obligan a poner la desviación típica entre paréntesis (DS). Ejemplo: Creatinina: 1.35 (0.28).

En líneas generales, vamos a utilizar en la inmensa mayoría de los casos tres tipos de descriptores estadísticos: Media y Desviación Estándar, Mediana y Rango Intercuartil (o percentil 25 y 75) y la Moda.

Media y Desviación Estándar

La media no es más que el punto de equilibrio o el centro de masas de la distribución de una variable. Es decir, si colocamos los valores a uno y otro lado de una balanza, la media sería el valor que mantiene la balanza en equilibrio. Su cálculo es sencillo:

$$\bar{X} = \frac{\sum_{i=1}^{i=n} Xi}{n}$$

Ejemplo: Valores de Creatinina: 1.2, 2.7, 5.4, 0.8, 1.6: $\bar{X} = \frac{\sum 1.2+2.7+5.4+0.8+1.6}{5} = 2.34$ mg/dl.

El valor obtenido puede ni siquiera aparecer en la distribución de los valores, es simplemente el valor que sirve como punto de equilibrio en la balanza.

Este valor aislado informa sobre cuál es el centro de masas de la distribución de los valores de una variable pero no informa sobre cómo están distribuidos esos valores, es decir, no informa sobre su dispersión. Los parámetros que informan sobre la dispersión de los valores indican cómo de cerca o de lejos están los datos alrededor de la media. Una distribución con mucha dispersión nos indica que los valores están muy alejados de la media. El parámetro que informa sobre la dispersión de los datos es la *Desviación Estándar*. Su cálculo es el siguiente:

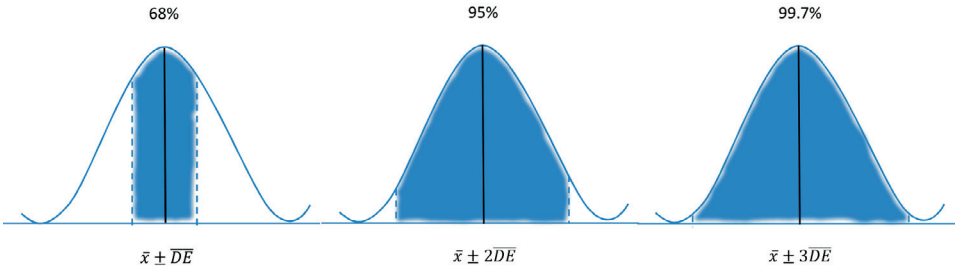
$$\overline{DE} = \sqrt{\frac{\sum_{i=1}^{i=n}(Xi-\bar{X})^2}{n-1}}$$

Ejemplo anterior: $DE=\sqrt{\frac{\sum_{i=1}^{i=n}(1.2-2.34)^2+(2.7-2.34)^2+(5.4-2.34)^2+(0.8-2.34)^2+(1.6-2.34)^2}{n-1}} = 1.85$

La ecuación anterior sin la raíz cuadrada es la fórmula de la *variancia*, parámetro que también informan sobre la dispersión. Por tanto la creatinina se expresaría como: 2.34 (1.85), NO como 2.34±1.85.

Sin embargo, aunque el valor de la media es un valor fácilmente comprensible, la desviación estándar es un término que difícilmente se entiende. Su uso sólo tiene

sentido cuando los datos se distribuyen de una manera simétrica alrededor de la media, es decir, como una distribución Normal. En este caso, el sumar o restar una vez, dos o tres veces la desviación estándar a la media nos da información de sobre qué valores se distribuyen el 68%, 95% y 99.7% de los datos tal y como se ilustra en los siguientes gráficos:



Por tanto, expresar los datos como media y desviación estándar sólo tendrá sentido cuando los valores se distribuyan como una distribución Normal y para ello primero tenemos que comprobar este supuesto.

Comprobación del supuesto de normalidad

La distribución Normal es una distribución cuya media es 0 y la desviación típica es 1. Es una distribución que se toma como referencia para calcular probabilidades en distribuciones de datos que se asemejen a la Normal. Existen tablas ya calculadas donde nos indican las probabilidades de encontrar valores superiores o inferiores de puntos predefinidos (denominados z); por ejemplo, la probabilidad de encontrar valores superiores a z= 1.51 es de 0.0655. Pero para poder usar la distribución Normal como referencia primero debemos comprobar si nuestra distribución de datos se asemeja a la normalidad. Generalmente se considera que una muestra sigue una distribución Normal si la "n" es suficientemente grande, y se considera grande una "n">30. Pero siempre interesa comprobar el supuesto de normalidad.

Como veremos a lo largo del curso, antes de la realización de algunos procedimientos estadísticos (como comparar medias con la *T-Student*), debemos previamente comprobar si los datos se distribuyen de manera Normal. Hoy en día, por suerte, los cálculos no se realizan a mano sino mediante programas estadísticos que nos permiten comprobar si nuestra muestra sigue una ley Normal o no. Para ello se utilizan las pruebas de Kolmogorov-Smirnov o de Shapiro-Wilk (esta última más recomendable cuando el tamaño de la muestra es pequeño, n<30) porque es más sensible para demostrar diferencias. Para muestras grandes Shapiro-Wilk detectará diferencias muy pequeñas aunque no sean relevantes y por eso es preferible la prueba de Kolmogorov-Smirnov). Ambas pueden realizarse con SPSS. El procedimiento es el siguiente para el caso de la PTH previa al trasplante de nuestra base de datos:

Analizar → Estadísticos Descriptivos → Explorar.

Seleccionamos la variable "PTH previa al trasplante" en el cuadro "Lista de Dependientes"; hacemos Click en el botón gráficos y seleccionamos la opción "Gráficos de Normalidad con Pruebas". La sintaxis es la siguiente:

EXAMINE VARIABLES=PTHPrevTx /PLOT NPPLOT /STATISTICS NONE
/CINTERVAL 95/MISSING LISTWISE /NOTOTAL.

El resultado es el siguiente:

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
PTH previa al trasplante (pg/ml)	,195	61	,000	,799	61	,000

a. Corrección de significación de Lilliefors

En nuestro caso, el tamaño de la muestra es de 60 casos (no es demasiado grande, pero sí mayor de 30). En cualquier caso, el valor tanto de Kolmogorov como de Shapiro es significativo ($p<0.001$), y por tanto nos está diciendo que los datos vulneran el supuesto de Normalidad (no siguen una ley Normal) y por tanto no podemos expresarlos como media y desviación estándar.

En algunos casos si los valores son muy extremos y su número no es muy importante, se puede describir la muestra con la Mediana Recortada un 5%. En este caso lo que hacemos es desechar el 5% de los valores que se encuentran a uno y otro lado de las colas izquierda y derecha de los gráficos anteriores. En cualquier caso, su uso no está muy extendido y generalmente es preferible utilizar la siguiente medida de centralización.

Mediana y Rango Intercuartil (o percentil 25 y 75)

En la distribución de los datos de una determinada variable es posible que haya valores muy extremos de tal forma que hagan que la media no sea la mejor manera de expresar los resultados. Por ejemplo, en una distribución de datos de creatinina con valores de 0.67, 0.86, 0.53, 0.97 y 24.2, el resultado de la media es 5.45. Este valor claramente no es representativo de esta distribución de datos. Tenemos 2 opciones, suprimirlo (no sería incorrecto, puesto que puede ser un error de laboratorio, una mala recogida de los datos o simplemente un valor que es encontrado en muy pocas ocasiones pero que por azar lo hemos obtenido) o utilizar otro estadístico.

Cuando tenemos pocos datos es fácil ver si tenemos valores extremos (no tenemos más que decirle a SPSS que nos muestre los valores mínimo y máximo, o mediante una exploración visual), pero cuando el tamaño de muestra es grande, la cosa se complica. Sin embargo, debido a los valores extremos, estas distribuciones muestrales no siguen una ley Normal, y por tanto lo único que tenemos que hacer es comprobar el supuesto de Normalidad. Si no lo cumplen debemos utilizar como medida centralizadora la Mediana.

Por tanto, si una muestra no sigue una ley Normal debemos utilizar la mediana para describir los datos. Tampoco pasaría nada sin aun siguiendo una ley Normal describiéramos

los datos con la mediana puesto que en este supuesto la media y la mediana coinciden. Si no queremos equivocarnos nunca podemos expresar los datos como mediana puesto que sirve tanto si sigue distribución Normal como si no.

El cálculo de la mediana es sencillo. Simplemente consiste en ordenar los datos de menor a mayor y ver qué valor es el que divide la muestra a la mitad.

En el ejemplo anterior sería: 0.53, 0.67, **0.86**, 0.97, 24.2. El valor 0.86 divide la muestra justo a la mitad y es más representativo que el valor de la media (5.45).

El cálculo del valor que divide la muestra a la mitad es: (n+1)/2.

En el ejemplo anterior sería (5+1)/2=3. El valor en la posición 3 divide la muestra en 2 partes iguales; para muestras impares es fácil.

Para muestras impares es un poco diferente. Si en el ejemplo anterior añadimos el valor 0.75, la distribución sería: 0.53, 0.67, 0.75, 0.86, 0.97, 24.2, y el valor obtenido con la fórmula anterior sería 3.5. En este caso se utiliza la media de los valores situados entre la posición 3 y 4, que en este caso sería (0.75+0.86)/2=0.81.

La mediana es el valor equivalente a la media para distribuciones Normales, pero necesitamos un valor (o valores) que den información sobre la dispersión de los datos como la desviación estándar. En este casos tenemos varias opciones: los percentiles 25 y 75 (o cuartiles 1 y 3, que son lo mismo), el rango intercuartil, o el mínimo y máximo. Los más utilizados son los percentiles 25 y 75 que son los valores que dividen en dos cada parte en que se ha dividido la muestra por la mediana; es decir, es hallar la mediana de cada parte. Su obtención es un poco laboriosa, y puesto que habitualmente vamos a tener muestras de tamaño mayor que el del ejemplo anterior y no vamos a calcular la mediana "a mano", vamos a utilizar el SPSS que nos da tanto el valor de la mediana como de los percentiles. El procedimiento es el siguiente para la variable "PTH previa al trasplante" de la base de datos:

Analizar → Estadísticos Descriptivos → Explorar

Seleccionamos la variable "PTH previa al trasplante" en el cuadro "Lista de Dependientes"; hacemos Click en el botón Estadísticos y seleccionamos la opción "Percentiles". La sintaxis es la siguiente:

```
EXAMINE VARIABLES=PTHPrevTx
/PLOT NONE
/PERCENTILES(5,10,25,50,75,90,95)HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

Las tablas resultantes son las siguientes:

Descriptivos

			Estadístico	Error estándar
PTH previa al trasplante (pg/ml)	Media		386,603	44,4705
	95% de intervalo de confianza para la media	Límite inferior	297,649	
		Límite superior	475,557	
	Media recortada al 5%		346,690	
	Mediana		247,000	
	Varianza		120635,012	
	Desviación estándar		347,3255	
	Mínimo		9,5	
	Máximo		1691,0	
	Rango		1681,5	
	Rango intercuartil		333,0	
	Asimetría		1,933	,306
	Curtosis		4,299	,604

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado (Definición 1)	PTH previa al trasplante (pg/ml)	68,570	82,160	156,350	247,000	489,350	889,800	1047,800
Bisagras de Tukey	PTH previa al trasplante (pg/ml)			157,000	247,000	480,700		

La primera tabla nos da información sobre la media, desviación estándar…, además de los valores mínimos y máximos donde podemos ver que hay valores extremos tanto por arriba como por abajo. En la segunda tabla aparecen los percentiles; vamos a tener en cuenta los valores obtenidos mediante las "Bisagras de Tukey". El percentil 50 corresponde a la mediana y los percentiles 25 y 75 al cuartil 1 y 3. Vemos la diferencia entre el valor de la mediana (247) y el de la media (386) debido a que no sigue una distribución Normal.

Por tanto para el caso de la PTH la forma correcta de expresarlo sería: PTH: 247 (P_{25}: 153.35, P_{75}: 489.35).

Entre los dos valores de los percentiles 25 y 75 se encuentran el 50% de los valores centrales de la variable PTH que nos informan de la dispersión de los mismos respecto a la medida de centralización (mediana).

****Concepto importante****

La tabla anterior sirve además para explicar otro concepto asociado a la media: vemos que aparecen la Desviación estándar (1ª columna) y Error estándar (2ª columna al lado de la media). *No son el mismo concepto*. La desviación estándar nos está dando información sobre cómo están distribuidos los datos de nuestra muestra, mientras que el error estándar nos está dando información sobre entre qué valores es posible que se encuentre la media de la población. Si a la media le sumamos y restamos 1.96 veces el error estándar obtendremos el intervalo de confianza del 95% en el que se encontrará la media poblacional. En este caso la media poblacional de PTH se situaría entre 299.4 y 473.8.

¿De qué debemos informar a la hora de comunicar resultados: de la desviación típica o del error estándar? Pues depende de nuestro objetivo: si estamos describiendo nuestra muestra usaríamos la desviación típica. Si por el contrario, lo que queremos hacer es dar información al lector sobre entre qué valores es posible que se sitúe la media en la población usaríamos el error estándar. Pero, ¿cómo lo hacemos? Es importante no usar los signos "±". Si

ponemos que los valores de PTH son de 386.6±44.5 el lector no va a saber si detrás del signo ± lo que hay es el error estándar o la desviación típica. Lo que vamos a hacer es lo siguiente:

- Si vamos a poner la desviación típica ponemos la media, y la desviación estándar entre ( ): PTH=386.6 (DE: 347.3).
- Si vamos a poner entre qué valores se sitúa la media de la población ponemos el intervalo de confianza al 95% (sumando y restando 1.96 veces el error estándar a la media): PTH=386.6 (IC95%: 299.4 – 473.8). Vemos que estos valores aparecen en el apartado "95% de IC para la media", aunque en este caso los valores se multiplican por 2 en lugar de por 1.96 (1.96≈2).

Moda

La Moda es un estadístico descriptivo muy poco utilizado, sin embargo, para determinadas variables es la mejor forma de describir su distribución. Se define como el valor que más veces se repite en una distribución.

Por ejemplo, en un estudio en el que se recoja el número de hijos de una paciente, no tiene ningún sentido que la media de hijos sea de 2.53. Es poco creíble que una persona tenga 2 hijos y un poco más de medio de otro. O tiene 2 o tiene 3. En este caso si el número de hijos que más frecuentemente tienen las pacientes es 2, éste sería el valor que mejor describiría la muestra.

Otros ejemplos parecidos lo tenemos con el número de embarazos, número de biopsias realizadas, número de ingresos de un paciente, número de incompatibilidades HLA, etc. Para acompañar a la Moda se puede dar información en forma de porcentajes sobre cuáles son los siguientes valores más frecuentes. En el caso anterior podríamos decir: el número de hijos que más frecuentemente tienen las pacientes es 2 (58% de las pacientes), seguidas de 1 hijo (20%) y de 4 hijos (12%).

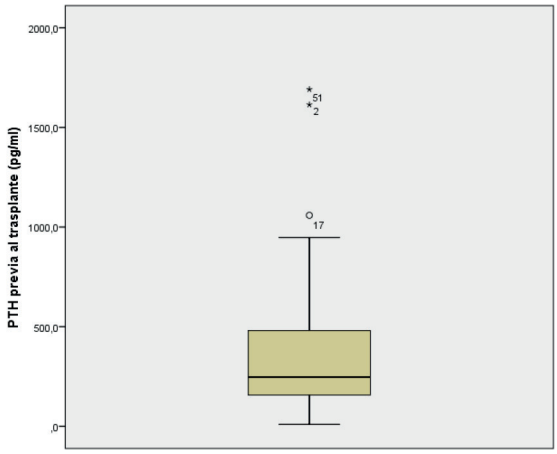
Mientras que sólo puede haber una media y una mediana, es posible que haya más de una moda si hay varios valores que se repiten con la misma frecuencia. Así por ejemplo, para el conjunto de valores: 1, 1, 2, 3, 3, 5, 6, 7, 8, 8, 9, los valores de la moda serían 1, 3 y 8.

REPRESENTACIÓN GRÁFICA DE VARIABLES CUANTITATIVAS

A la hora de describir las variables cuantitativas de un determinado estudio, aparte de los valores concretos que hemos visto en los apartados anteriores, generalmente se suele acompañar una representación gráfica de los mismos. Los gráficos van a aportar información visual de los datos y en algunos casos incluso pueden sustituir algunas tablas a la hora de escribir un artículo.

Hay diversos gráficos para representar variables cuantitativas (Histogramas, Polígono de frecuencias), pero la representación gráfica más correcta es mediante un Diagrama de Cajas (BoxPlot).

La representación gráfica de la PTH de nuestro estudio sería la siguiente:



En este gráfico podemos ver representados los siguientes datos:

- Una caja central que contiene el 50% de los datos; por tanto la cara de arriba de la caja corresponde al P_{75} y la cara de abajo al P_{25}. La altura de la caja nos va a dar información de la distribución del 50% de los datos centrales.
- Una línea en el interior de la caja que nos informa sobre la Mediana.
- Unas patillas a ambos lados de la caja que representan los valores mínimo y máximo que no se consideran "anormales".
- Informa sobre la presencia de valores muy extremos (representados con un círculo), que son aquellos valores que están por encima del P75 o por debajo del P_{25} más de 1.5 veces el valor del rango intercuartil ($P_{75} - P_{25}$).
- Informa sobre valores alejados, representados por un asterisco, que son aquellos valores que están por encima del P_{75} o por debajo del P_{25} más de 3 veces el valor del rango intercuartil ($P_{75} - P_{25}$).

En este caso vemos que el valor del caso número 17 es un valor extremo, mientras que los valores de los casos números 2 y 51 son valores muy alejados.

En caso de tratarse de una distribución normal, el diagrama sería simétrico y el eje de simetría pasaría por la mediana (puesto que en este caso la media coincidiría con la mediana).

DESCRIPCIÓN DE VARIABLES CUALITATIVAS (CATEGÓRICAS)

La descripción de variables categóricas se acostumbra a realizar mediante el número de casos de una determinada categoría acompañada del porcentaje que representa del total de la variable.

Por ejemplo para la variable Sexo de nuestro ejemplo obtenemos el número para cada categoría y el porcentaje del mismo mediante un análisis de frecuencias. El procedimiento sería:

Analizar → Estadísticos Descriptivos → Frecuencias. Seleccionamos la variable Sexo y obtenemos el siguiente cuadro de diálogo:

Debemos tener marcada la opción "Mostrar tablas de frecuencias" para que nos represente la tabla. Este procedimiento también sirve para variables cuantitativas (pero debemos NO tener seleccionada la opción "Mostrar tablas de frecuencias" pues la tabla obtenida tendría un gran tamaño ya que nos representaría cada uno de los valores con su porcentaje). En el botón de la derecha "Estadísticos" podemos seleccionar la opción de Media, Mediana, Desviación Estándar…. En el botón "Gráficos podemos seleccionar las opciones de *Gráficos de Barras* (para variables categóricas o cuantitativas discretas), *Gráficos de sectores* (el famoso gráfico de quesitos para variables categóricas) o *Histogramas* (con curva Normal para variables cuantitativas). A su vez podemos seleccionar que represente en el gráfico el recuento o el porcentaje. En el botón "Formato" podemos ordenar la tabla en orden ascendente o descendente.

En nuestro caso la tabla sería:

Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Hombre	40	65,6	65,6	65,6
	Mujer	21	34,4	34,4	100,0
	Total	61	100,0	100,0	

Vemos que aparecen 3 porcentajes (Porcentaje, Porcentaje válido y Porcentaje acumulado). La columna "Porcentaje" no tiene en cuenta los valores perdidos mientras que la columna "Porcentaje válido" sí los tiene en cuenta y por tanto los datos en los que nos vamos a fijar son los de esta columna. En este caso como no hay ningún valor perdido coinciden ambas columnas. La última columna simplemente va sumando los porcentajes encontrados en las sucesivas categorías (lógicamente en la última categoría siempre va a ser 100%). La primera columna informa sobre el número de casos de cada categoría.

Para la variable "Diagnósticos de IRC" la tabla sería:

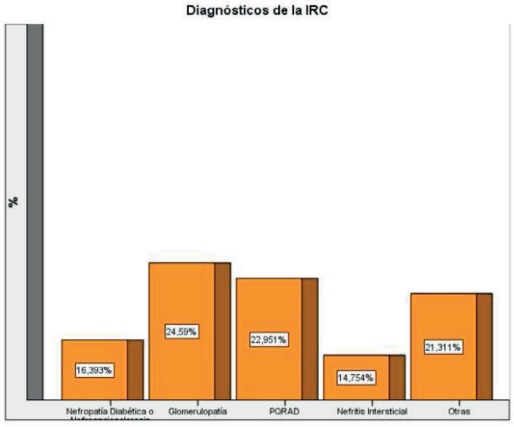
Diagnósticos de la IRC

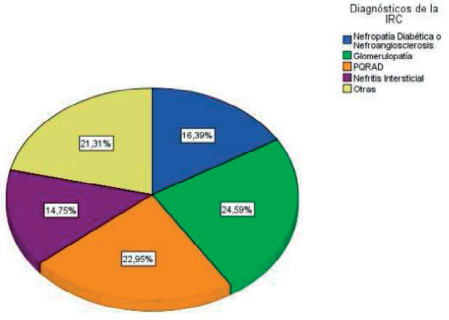
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Nefropatía Diabética o Nefroangiosclerosis	10	16,4	16,4	16,4
	Glomerulopatía	15	24,6	24,6	41,0
	PQRAD	14	23,0	23,0	63,9
	Nefritis Intersticial	9	14,8	14,8	78,7
	Otras	13	21,3	21,3	100,0
	Total	61	100,0	100,0	

REPRESENTACIÓN GRÁFICA DE VARIABLES CUALITATIVAS

Las variables categóricas se acostumbran a representar gráficamente mediante diagramas de barras o los diagramas de sectores (diagrama de "quesitos").

La representación gráfica del diagnóstico de IRC de nuestro estudio sería el siguiente:





Modificando los ajustes de los gráficos con un doble Clic sobre el mismo podemos cambiar el aspecto del mismo (cambiar color, añadir títulos, que aparezcan los porcentajes o recuentos, darles volumen,…).

ANÁLISIS ESTADÍSTICO DE LOS DATOS

Una vez descrita nuestra muestra podemos pasar a realizar análisis estadísticos en la misma. El análisis estadístico trata de buscar "causalidades" entre las variables del estudio, pero para ello previamente debemos haber comprobado que no hay "casualidades". Es decir, si queremos decir que tal variable produce tal efecto en otra, previamente debemos haber demostrado que no es fruto del azar.

Generalmente, salvo en los estudios de Correlación, vamos a querer ver si una determinada variable (que denominaremos variable "Predictora" o "Independiente") produce un determinado efecto sobre otra variable (que denominaremos "variable dependiente" puesto que sus valores van a depender de los que tome la variable predictora). Por ejemplo, en nuestra base de datos vamos a querer comprobar si la Acidosis Metabólica se asocia a mayor riesgo de Osteoporosis. En este caso, la Acidosis Metabólica será la variable Independiente y la Osteoporosis la variable Dependiente.

Dependiendo de si la variable independiente y dependiente es cuantitativa o categórica vamos a realizar análisis estadísticos distintos.

- Variable dependiente cuantitativa:
  - Variable independiente cuantitativa: análisis de Correlación.
  - Variable independiente categórica:

    2 categorías: Pruebas T-Student o U de Mann-Whitney según el caso.

    Más de 2 categorías: Análisis de la Variancia (ANOVA).
- Variable dependiente categórica:
  - Variable independiente cuantitativa: Regresión Logística.
  - Variable independiente categórica: Prueba de Chi-cuadrado ($\chi 2$) o prueba exacta de Fisher según el caso.

Independientemente de cómo sea la variable independiente, existen métodos estadísticos más potentes que permiten además introducir otras variables de ajustes; estos son los métodos de Regresión. Tenemos los siguientes tipos:

- Regresión Lineal: Variable dependiente cuantitativa.

- Regresión Logística Binaria: Variable dependiente categórica binaria. Si son varias categorías se utiliza la Regresión Logística Multinomial.
- Regresión de Cox: Variable dependiente binaria pero donde además se incluye el factor tiempo. Permite además analizar curvas de supervivencia.

Antes de pasar a describir los distintos procedimientos estadísticos vamos a repasar unas nociones básicas sobre cómo se comprueban hipótesis.

COMPROBACIÓN DE HIPÓTESIS

Previo a la realización de cualquier estudio vamos a elaborar una hipótesis que trataremos de demostrar tras realizar el análisis estadístico de los datos que hemos recogido. Pero, ¿cómo se comprueba una hipótesis? Estamos habituados a ver en las publicaciones científicas valores de significación "p" e intervalos de confianza, pero ¿esto realmente qué quiere decir?

Por ejemplo, supongamos que en nuestro estudio queremos comprobar si los valores de bicarbonato son más bajos en los pacientes con IRC que en los pacientes sin IRC, y supongamos que en nuestra provincia hemos extraído una muestra de 60 pacientes con IRC y presentan valores de bicarbonato de media 17 mmol/l y desviación estándar 3.5 mmol/l. Supongamos además que en la población española (población de referencia) se ha medido el nivel de bicarbonato a todas las personas sin IRC y presentan una media de 24 mmol/l y desviación estándar de 2.5 mmol/l. ¿Cómo podemos saber si nuestra media de bicarbonato de 17 es un valor realmente bajo o por el contrario es un valor que podríamos encontrar en pacientes sin IRC? Bastaría con comprobar la probabilidad de encontrar un valor de 17 mmol/l en pacientes sin IRC. Para ello tendríamos que tomar como distribución de referencia la distribución de las medias de todas las muestras con 60 pacientes extraídas de la población cuya media es 24 y comprobar la probabilidad de poder encontrar un valor inferior a 17 en esta distribución. Dado que procede de muestra grande y por tanto con distribución Normal, podemos utilizarla como referencia y buscar en las Tablas de Normalidad la probabilidad de encontrar un valor "z" que corresponda a un valor inferior a 17. En este caso el valor de $p<0.001$. ¿Qué significa esta p? Simplemente quiere decir que la probabilidad de encontrar un valor de bicarbonato de 17 en pacientes sin IRC es de 1 por mil. Como es una probabilidad muy baja podemos decir que los valores de bicarbonato en los pacientes con IRC son menores que en pacientes sin IRC. Pero ¡¡¡Ojo!!! También es posible encontrar esos valores en pacientes sin IRC aunque la probabilidad sea muy baja. ¿Qué punto de corte elegimos para asegurar que una probabilidad es baja? Pues por consenso se estableció poner el corte en una $p=0.05$ (que como podemos ver no es un valor realmente bajo; en nuestro ejemplo significaría que la probabilidad de encontrar valores de bicarbonato de 17 en pacientes sin IRC es del 5% (realmente es un valor importante). Si existiera una casa de juego de azar en el que la probabilidad de ganar fuera del 5% es posible que la casa de apuestas se arruinara.

En el supuesto anterior hemos establecido que un valor sea inferior que otro, pero generalmente lo que vamos a tratar de comprobar es simplemente si los valores son diferentes (mayores o menores). En nuestro ejemplo sería comprobar si los valores de bicarbonato de los

pacientes con IRC son diferentes (ni mayores ni menores) que los de los pacientes sin IRC. El mecanismo de comprobación es similar al anterior salvo porque se va a utilizar la distribución de las medias de las diferencias como distribución de referencia. En efecto, si no existe diferencias entre los niveles de bicarbonato de una y otra muestra la diferencia de sus medias será 0. Extrayendo todos los pares de muestras de igual tamaño a la nuestra de la población, cuya media va a valer 0, nuestro objetivo va a ser comprobar la probabilidad de encontrar una diferencia de 7 mmol/l, que es lo que hemos encontrado. De nuevo acudiremos a las tablas de normalidad para calcular la probabilidad. En este caso si el valor es p=0.01 la conclusión será que la probabilidad de que las medias de ambas muestras (con y sin IRC) sean iguales es igual al 1% (pero existe esa pequeña posibilidad).

Con las conclusiones anteriores tan sólo vamos a poder decir si las diferencias observadas son significativas o no, pero no si esa significación tiene relevancia. Para ayudar a ello se utilizan los intervalos de confianza. Un intervalo de confianza no es más que los límites sobre los que se va a encontrar un parámetro en cuestión con una probabilidad prefijada. En nuestro ejemplo si tenemos un intervalo de confianza del 95% (IC95%) de las diferencias de bicarbonato entre ambos grupos entre 4 y 9, significa que la media de las diferencias se va a situar entre ambos valores con una confianza del 95%. Ahora nosotros podemos determinar si ese intervalo tiene relevancia o no. Si en este caso el intervalo incluye el valor 0 significa que no hay diferencias entre ambos grupos y por tanto tampoco será significativo. Podemos establecer un punto a priori para determinar la relevancia. Por ejemplo, imaginemos que estudios previos han determinado que sólo diferencias de bicarbonato superiores a 10 mmol/l tienen repercusión en la supervivencia de los pacientes; pues en este caso, las diferencias son significativas pero como el intervalo no incluye el valor 10, podemos decir que no son clínicamente relevantes.

Por otra parte vemos en determinados estudios que se calcula la potencia del mismo, y el tamaño de muestra. ¿Qué significa esto? Son términos extraídos de las pruebas de hipótesis. Las pruebas de hipótesis establecen previo a realizar un estudio un valor de diferencias que se va a considerar como clínicamente relevante, estableciendo además unos parámetros previos para considerar que no son debidos al azar (valores α y β). Consideran dos hipótesis previamente a la realización del estudio: hipótesis nula (H_0) que establece que no hay diferencias, e hipótesis alternativa (H_1) que establece que sí las hay. Con los resultados se pueden cometer 2 errores: Error Tipo I (α) que consiste en decir que existen diferencias cuando realmente no las hay. Y Error Tipo II (β) que consiste en decir que no hay diferencias cuando realmente sí las hay. Prefijando los términos anteriores (valor de la diferencia que se va a considerar como relevante, valor α –generalmente del 5%–, y valor β –generalmente del 20% considerado como potencia del estudio–) vamos a poder calcular además el número de pacientes necesarios para su comprobación. Esta metodología de comprobación de hipótesis suele utilizarse para diseñar estudios prospectivos como los Ensayos Clínicos.

Un dato a tener en cuenta es el tamaño de la muestra. Cuanta más pequeña sea la diferencia que queramos demostrar mayor debe ser el tamaño de la muestra. Por tanto, para demostrar que una pequeña diferencia es estadísticamente significativa lo único que tenemos que hacer es introducir muchos pacientes en el estudio. Esto no significa que esa

pequeña diferencia sea clínicamente relevante. Estamos acostumbrados a que nos presenten fármacos, como los antihipertensivos, donde nos enseñan gráficos, porcentajes, etc., donde nos dicen que el fármaco en cuestión consiguió disminuir de manera significativa la tensión arterial con valores de "p" muy bajos; si analizamos esos datos con detenimiento podemos comprobar en muchas ocasiones que a lo mejor el descenso de la tensión arterial fue sólo de 2 mmHg realizado en 15.000 pacientes. Efectivamente será significativo, pero la relevancia clínica de un descenso de 2 mmHg de la tensión arterial habrá que ponerla en cuestión.

CÁLCULO DEL TAMAÑO MUESTRAL

En determinados estudios, como los Ensayos Clínicos donde vamos a invertir mucho tiempo y posiblemente mucho dinero, nos va a interesar determinar de una manera aproximada, cuál debe ser el tamaño de la muestra antes de comenzar el estudio. De esta manera nos aseguraremos que las diferencias que intentamos demostrar sean estadísticamente significativas (error alfa) y con una potencia suficiente (error beta). Como vemos, vamos a predeterminar cual es el error alfa y la potencia del estudio que deseemos. Además, deberemos especificar cuál es la diferencia que consideramos como clínicamente relevante.

Como se ha comentado anteriormente, si queremos demostrar una pequeña diferencia, lo único que tenemos que hacer es aumentar el tamaño de la muestra, por tanto, la utilización de estas fórmulas tratan de garantizar un tamaño de muestra suficiente que permita demostrar esas diferencias. El valor de la diferencia que se considera como clínicamente relevante es establecido por el investigador. De esta manera se diseña un tamaño de muestra "a la carta".

Existen varias calculadoras en internet que nos permiten calcular este tamaño de muestra, pero debemos tener cuidado de cual escoger. La página web: www.statsol.ie, con el programa *nQuery Advisor* es la más recomendada. No obstante, se pueden calcular a mano sin ningún tipo de problemas, tan sólo vamos a necesitar algunos parámetros a priori.

Estas fórmulas van a depender de si comparamos medias o proporciones. Además dependerá de si la prueba es unilateral o bilateral y de si la población es finita o infinita. Generalmente vamos a realizar estudios con pruebas bilaterales en poblaciones infinitas.

Las fórmulas son:

- Para comparar medias:

$$n = \frac{2\sigma^2\left(Z_\alpha + Z_{\beta/2}\right)^2}{\delta^2}$$

Dónde:

σ: Desviación estándar, suponiendo que $\sigma_1=\sigma_2=\sigma$.

δ: Diferencia que consideramos clínicamente relevante.

• Para comparar proporciones:

$$n = \frac{2\pi(1-\pi)\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\delta^2}$$

Dónde:

μ: Proporción, siendo $\pi=(\pi_1+\pi_2)/2$.

δ: Diferencia que consideramos clínicamente relevante.

Los valores de Z_α y Z_β los vemos en la siguiente tabla:

		Error α				
		0.10 (10%)	0.05 (5%)	0.025 (2.5%)	0.01 (1%)	0.005 (0.5%)
Error β	0.20 (20%)	4.5079	6.1826	7.8489	10.0360	11.6790
	0.15 (15%)	5.3731	7.1893	8.9784	11.3083	13.0484
	0.10 (10%)	6.5695	8.5638	10.5074	13.0169	14.8794
	0.05 (5%)	8.5638	10.8222	12.9947	15.7704	17.8142
	0.025 (2.5%)	10.5074	12.9947	15.3658	18.3725	20.5734

Como podemos ver, hay parámetros que se necesitan antes de realizar el propio estudio, como la desviación estándar o las proporciones. Si aún no hemos hecho el estudio, ¿cómo vamos a saberlo? Pues se deben revisar otras publicaciones en las que se describan estos valores o en estudios pilotos previo que hayamos hecho.

El valor de la "n" obtenido es el tamaño de cada grupo, suponiendo que ambos grupos van a tener el mismo tamaño.

Ejemplo 1: se quiere saber el tamaño de muestra necesario para comparar la eficacia de 2 fármacos reductores de los niveles de PTH en pacientes con IRC, con un riesgo α del 5% y β del 10%. Se sabe por otros estudios que la desviación estándar de la PTH en pacientes con IRC es de 115 pg/ml, y que se considera una diferencia clínicamente relevante de 150 pg/ml. El tamaño sería:

El valor de $Z_\alpha + Z_{\beta/2}$ sería para un α igual a 0.05 y un $\beta/2$ igual a 0.05 (10/2): 10.8222.

$$n = \frac{2 \times 115^2 (10.8222)^2}{150^2} = 137.68 = 138$$

Se redondea al entero superior. En este ejemplo el tamaño sería de 138 sujetos por grupo.

Ejemplo 2: queremos hacer un estudio para ver si la corrección de la acidosis metabólica reduce el riesgo de osteoporosis al año del trasplante, con un riesgo α del 5% y β del 10%. Sabemos por otros estudios que la proporción de osteoporosis al año es del 20%, y se considera una diferencia clínicamente relevante del 15% (para una conseguir una proporción del 5% similar al de la población normal no acidótica).

$\pi=(0.20 + 0.05)/2=0.125$

El valor de $Z_{\alpha/2} + Z_{\beta}$ sería para un $\alpha/2$ (5/2) igual a 0.025 y un β igual a 0.10: 10.5074.

$$n = \frac{2 \times 0.125(1 - 0.125) \times 10.5074^2}{0.15^2} = 1226.7 = 1227$$

En este otro ejemplo el tamaño por grupo sería de 1227 sujetos.

Cuando no se conoce por otros estudios el valor de π, se utiliza el valor 0.5 (que implica igual proporción en ambos grupos y aporta el tamaño de muestra más alto).

ANÁLISIS ESTADÍSTICO BÁSICO CON VARIABLES CUANTITATIVAS

En este apartado vamos a estudiar tres aspectos de las variables cuantitativas:

- Relacionar 2 variables cuantitativas entre sí. Ejemplo: ver la relación entre la creatinina y el bicarbonato.
- Ver la diferencia de una variable cuantitativa en 2 grupos: Ejemplo: ver la diferencia en las cifras de bicarbonato entre los pacientes con o sin insuficiencia renal crónica.
- Ver la diferencia de una variable cuantitativa en más de 2 grupos. Ejemplo: ver la diferencia en las cifras de bicarbonato según los estadios de función renal.

RELACIÓN ENTRE 2 VARIABLES CUANTITATIVAS

Imaginemos que queremos ver si hay relación entre las cifras de creatinina y de bicarbonato, ambas como variables cuantitativas. Realmente no queremos ver si la creatinina aumenta o disminuye como consecuencia del bicarbonato, o viceversa; es decir, no tenemos variable independiente ni dependiente. Ambas variables juegan un papel simétrico y tan sólo queremos ver si hay asociación entre ellas. Esto es lo que llamamos Estudio de Correlación.

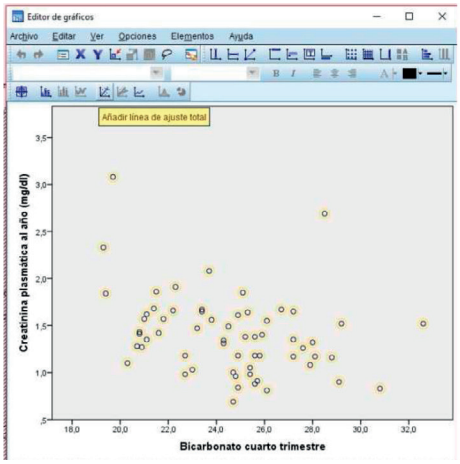
Básicamente consiste en representar una nube de puntos con los valores de cada una de las variables y ajustar una recta a esa nube de puntos. Si la recta es horizontal significa que no hay asociación (puesto que los valores de una variable no se modifican con los cambios de los valores de la otra variable) y si la recta tiene inclinación positiva o negativa significa que sí hay asociación.

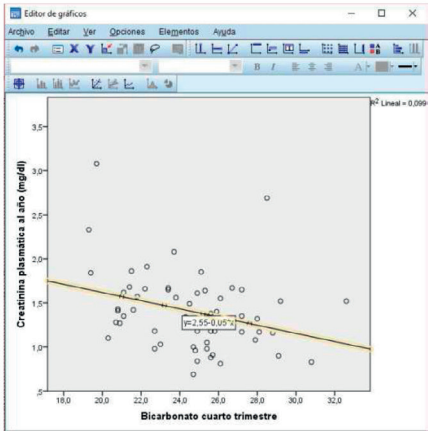
La asociación se valora mediante el Coeficiente de Correlación, que toma valores entre - 1 (asociación lineal negativa completa) y +1 (asociación lineal positiva completa). El valor 0 indica ausencia de asociación (recta horizontal).

La significación estadística se efectúa mediante el Coeficiente de Correlación de Pearson, pero para ello se necesita que la distribución de ambas variables siga una distribución normal. De no ser así, se debe utilizar el Coeficiente de Spearman.

Veamos la relación entre creatinina y bicarbonato al año en nuestro estudio. Primero vamos a dibujar el gráfico de puntos para ver la distribución de la nube. Seleccionamos:

Gráficos → Generador de gráficos. Y elegimos el gráfico de puntos, arrastrando a los ejes X e Y las variables. Tras generarlo, hacemos doble Clic sobre el mismo y seleccionamos la opción "Añadir línea de ajuste total". El resultado es el siguiente:

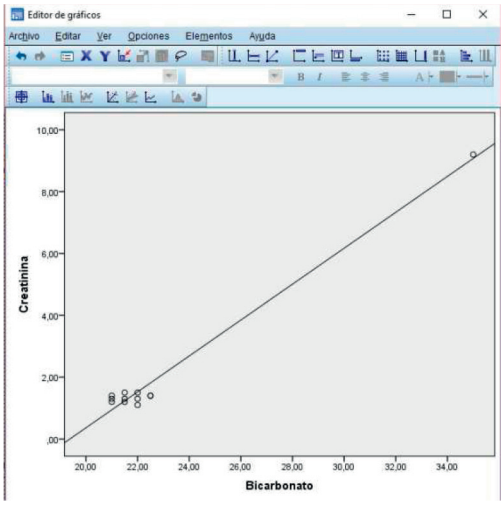




Podemos ver que la recta resultante tiene tendencia negativa, es decir, a medida que aumentan las cifras de bicarbonato disminuyen las cifras de creatinina. Arriba a la derecha aparece el valor R^2 Lineal = 0.099 que no es más que el cuadrado del coeficiente de correlación de Pearson "R". En el centro del cuadro aparece el valor de la ecuación de la recta Y= a +bX, que en este caso vale Y=2.55-0.05*X. El valor de la constante "a" es el punto de corte sobre el eje de ordenadas cuando la abscisa vale 0; es decir, teóricamente cuando el valor del bicarbonato fuese 0, la creatinina valdría 2.55. Esta ecuación podemos hacer que no aparezca editando el gráfico y deseleccionando la opción "Adjuntar etiqueta a la línea". Debemos tener en cuenta que la recta resultante es la hallada para estos valores de creatinina y de bicarbonato que hemos recogido, que vemos que están comprendidos entre más o menos 3.5 y 32 respectivamente, pero no podemos asegurar que la fórmula satisfaga las características de los pacientes con valores fuera de los rangos recogidos. No sería esperable que para un paciente con 0 de bicarbonato las cifras de creatinina fuera tan sólo de 2.55.

Para la estimación de la recta anterior lo que se tiene en cuenta es que la distancia de cada punto a la recta sea la menor posible. Dibuja la recta de manera que el cuadrado de la distancia de cada punto para cada valor de bicarbonato sea la menor posible.

Antes de decidir si existe asociación o no, debemos ver cómo es la nube de puntos. Imaginemos que los valores de creatinina para un valor de bicarbonato entre 20 y 22 estuvieran en torno a 1.5 y tuviésemos un solo valor de creatinina de 9 para un valor de bicarbonato de 35 como en el siguiente gráfico:



Vemos que la recta tiene una tendencia claramente positiva, y a simple vista parece haber una gran asociación, sin embargo esta recta es a expensas de un único valor muy alejado y no parece amoldarse al resto de valores de la nube de puntos que parecen no tener asociación pues a simple vista parece que seguirían una línea horizontal. Este valor probablemente tendremos que eliminarlo de nuestra base de datos (o corregirlo si ha sido un error de recogida de datos) puesto que nos va a dar una ecuación de la recta y una asociación que no es aplicable para la mayoría del resto de valores. Este hecho debemos tenerlo muy en cuenta cuando hablemos de diagnósticos de modelos de regresión, cuya base fundamental es similar a la correlación.

El siguiente paso es valorar si esta asociación es significativa o no. En primer lugar debemos comprobar si las variables siguen o no una distribución Normal (mediante Shapiro-Wills o Kolmogorov). Si siguen distribución Normal se utiliza el coeficiente de correlación de Pearson. Si no sigue una distribución Normal debemos utilizar el de Spearman, de la siguiente manera:

Analizar → Correlaciones → Bivariadas. Seleccionamos las variables de estudio y marcamos los coeficientes de Pearson y de Spearman. La sintaxis es:

CORRELATIONS /VARIABLES=Cr4 HCO34

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE. NONPAR CORR

/VARIABLES=Cr4 HCO34

/PRINT=SPEARMAN TWOTAIL NOSIG

/MISSING=PAIRWISE.

Y la tabla resultante es:

Correlaciones

		Cr4	HCO34
Cr4	Correlación de Pearson	1	-,314*
	Sig. (bilateral)		,014
	N	61	61
HCO34	Correlación de Pearson	-,314*	1
	Sig. (bilateral)	,014	
	N	61	61

*. La correlación es significativa en el nivel 0,05 (bilateral).

Correlaciones

			Cr4	HCO34
Rho de Spearman	Cr4	Coeficiente de correlación	1,000	-,343**
		Sig. (bilateral)	.	,007
		N	61	61
	HCO34	Coeficiente de correlación	-,343**	1,000
		Sig. (bilateral)	,007	.
		N	61	61

**. La correlación es significativa en el nivel 0,01 (bilateral).

Vemos que el coeficiente de correlación de Pearson es significativo con p=0.014. Si no sigue distribución Normal, también es significativo, con Rho de Spearman de 0.007. Es decir, hay correlación entre valor de bicarbonato y creatinina. Esta correlación es negativa puesto que el valor de R=-0.314, es decir, a medida que aumenta una disminuye la otra. Si elevamos al cuadrado este valor obtendremos el valor del R2 que aparece en el gráfico 0.099; significa que ambas variables tienen un 9.9% de variabilidad común. R cuanto más se acerque a 1 o a -1 mejor, puesto que al elevarlo al cuadrado sería también 1, es decir, las variables tienen un 100% de variabilidad común.

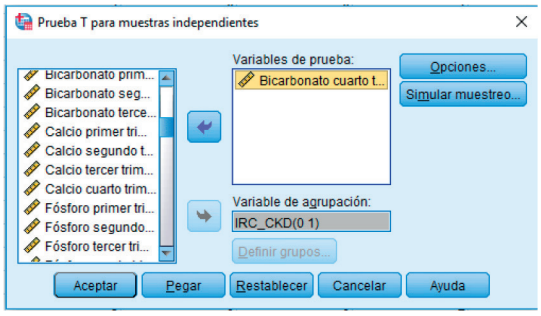
COMPARACIÓN DE UNA VARIABLE CUANTITATIVA EN 2 GRUPOS DISTINTOS.

Vamos a ver si los valores de una determinada variable cuantitativa son diferentes en 2 grupos distintos. Como ejemplo queremos ver si los valores de bicarbonato son diferentes en los pacientes con y sin IRC.

Para ello se acostumbra a comprobar si las medias de bicarbonato de ambos grupos son diferentes. Si las medias son iguales no habrá diferencias (la diferencia de medias=0). Para el cálculo de la significación estadística utilizaremos el estadístico T-Student. Pero previo a ello debemos comprobar de nuevo que las variables en cada categoría de estudio siguen una distribución Normal pues de lo contrario no se puede aplicar este estadístico.

Una vez comprobado si siguen distribución Normal el procedimiento es el siguiente:

Analizar → Comparar medias → Prueba T para muestras independientes. Se nos abrirá un cuadro de diálogo como el siguiente:



Seleccionamos la variable cuantitativa a comparar, en este caso Acidosis al año (HCO34) y en los grupos en los que lo vamos a comparar, en este caso la presencia o no de IRC;

seleccionaremos la "variable de agrupación" IRC-CKD y Definiremos los grupos con los valores que le hemos dado (en este caso 0 y 1 que correspondían a sin y con IRC).

Ejecutando la sintaxis el resultado es el siguiente:

Estadísticas de grupo

	IRC_CKD	N	Media	Desviación estándar	Media de error estándar
HCO34	No	27	25,170	2,4542	,4723
	Sí	34	23,997	3,2040	,5495

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
HCO34	Se asumen varianzas iguales	3,648	,061	1,571	59	,122	1,1733	,7469	-,3213	2,6679
	No se asumen varianzas iguales			1,619	58,940	,111	1,1733	,7246	-,2766	2,6232

Una primera tabla donde se especifica la media y su correspondiente desviación estándar en cada grupo. Podemos ver además cuantos pacientes se han tenido en cuenta en cada grupo para calcular los datos (puesto que los pacientes con valores perdidos no se tienen en cuenta). Vemos que la media de bicarbonato para los pacientes con IRC es menor que para los pacientes sin IRC. La segunda tabla nos indica la significación de esta diferencia y entre qué valores se va a situar esa diferencia en la población. Debemos tener en cuenta una cosa en esta tabla: vemos que hay una columna que pone "Prueba de Levene de igualdad de varianzas". Primero debemos ver si la varianza en ambos grupos es igual o no, puesto que de no serlo la significación del análisis no es la misma. Si las varianzas son iguales, la prueba de Levene no será significativa y tendrá una $p>0.05$ (es decir, se acepta la hipótesis H_0); por el contrario, si las varianzas no son iguales (H_1) el valor será significativo ($p<0.05$). Si se asume que las varianzas son iguales, se toma el valor de significación para la prueba t que aparece en la primera fila, y si no son iguales las varianzas se toma el valor de la segunda fila. En este caso, el valor de la prueba de Levene tiene una significación de 0.061 (no significativo), por tanto se asumen que las varianzas son iguales y por tanto el valor de significación de la prueba t es $p=0.122$ (no significativo). Significado: aunque los valores de bicarbonato en los pacientes con IRC son inferiores al de los pacientes sin IRC, esta diferencia no es significativa. Como vemos en la columna "diferencias de medias", los pacientes sin IRC tienen de media 1.1733 mEq/l de bicarbonato más que los pacientes con IRC; esta diferencia oscila entre -0.3213 y 2.6679 mEq/l. Dado que este intervalo engloba el valor 0 ya se asume que la diferencia no es significativa (sin necesidad incluso de ningún valor p). Para mostrar estos datos en un artículo podemos decir: 1) que los niveles de bicarbonato en ambos grupos son iguales; 2) que no hay diferencias significativas en los valores de bicarbonato entre ambos grupos; o 3): que la diferencia de los valores de bicarbonato en los pacientes sin IRC y con IRC es de 1.1733 mEq/l a favor de los pacientes sin IRC con un IC95% entre -0.3213 a 2.6679, $p=0.122$.

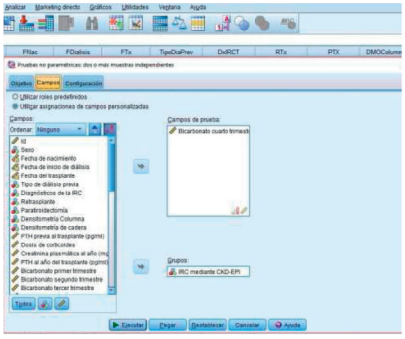
Si la distribución de los datos en ambos grupos no sigue una distribución Normal, no podemos usar la prueba t-Student. En este caso debemos usar métodos no paramétricos, siendo la prueba más utilizada la U de Mann-Whitney. Sin entrar en detalles en su desarrollo,

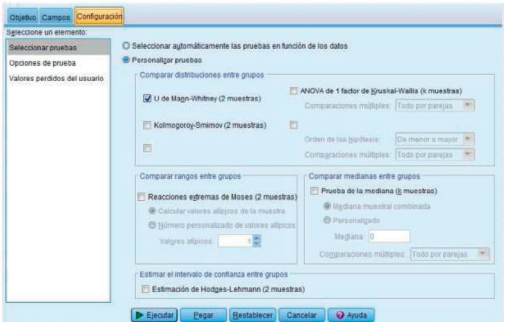
simplemente decir que se trata en ordenar todas las diferencias de observaciones de menor a mayor sin tener en cuenta el signo en cada grupo, asignándole a cada observación un número de orden. Dado que en el ejemplo anterior sí sigue una distribución Normal, este procedimiento sólo se va a utilizar como ejemplo. El procedimiento con SPSS sería el siguiente:

Analizar → Pruebas no paramétricas → Muestras independientes. La sintaxis es la siguiente:

```
NPTESTS
/INDEPENDENT TEST (HCO34) GROUP (IRC_CKD) MANN_WHITNEY
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.
```

Se nos abrirá el siguiente cuadro de diálogo:





En la pestaña "Objetivo" marcamos la opción "personalizar análisis". En la pestaña "campos" seleccionamos los campos a comparar, y en la pestaña "configuración" marcamos la opción "personalizar pruebas" y posteriormente la opción "U de Mann- Whitney (2 muestras). Tras ejecutar la sintaxis, el resultado es el siguiente:

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de Bicarbonato cuarto trimestre es la misma entre las categorías de IRC mediante CKD-EPI.	Prueba U de Mann-Whitney para muestras independientes	,093	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

Vemos que la significación es p=0.093 y que incluso nos indica el programa que conservemos la hipótesis nula donde la "distribución de Bicarbonato cuarto trimestre es la misma entre las categorías de IRC mediante CKD-EPI".

Si hacemos doble Clic sobre esta tabla, se nos abrirá una ventana en la que aparecerá el resumen de la prueba con los valores de la U de Mann-Whitney, prueba de los signos de Wilcoxon, etc.

Es posible que en determinadas versiones de SPSS no funcione este cuadro de diálogo. En este caso el procedimiento es el siguiente:

Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos → 2 muestras independientes. La sintaxis es:

NPAR TESTS /M-W= HCO34 BY IRC_CKD(0 1)
/MISSING ANALYSIS.

Se nos abrirá el siguiente cuadro de diálogo, donde seleccionaremos la variable a comparar en función de los grupos y marcaremos la opción "U de Mann-Whitney".

Tras ejecutar la sintaxis, el resultado es el siguiente:

Estadísticos de prueba[a]

	HCO34
U de Mann-Whitney	343,500
W de Wilcoxon	938,500
Z	-1,678
Sig. asintótica (bilateral)	,093

a. Variable de agrupación: IRC_CKD

Vemos que el resultado es igual que con el procedimiento anterior.

En este caso (si no siguen distribución Normal) para describir los datos en un estudio, no debemos poner la media (DE), sino la mediana y los percentiles 25 y 75, informando de que la prueba de significación se ha realizado con la U de Mann-Whitney.

En determinados estudios, no vamos a tener grupos que comparar, sino que vamos a tener una observación de un parámetro y queremos ver si nuestra observación difiere de lo que ya hay descrito en la literatura. Por ejemplo, imaginemos que tan sólo hemos recogido los datos de los valores de bicarbonato sólo en los pacientes con IRC, y queremos ver si son diferentes respecto a los valores de los pacientes sin IRC, y hemos leído en un artículo que alguien ha recogido esos valores y su media es de 24 mEq/L. ¿nuestros valores son o no diferentes a los descritos? Es decir, vamos a comparar una media observada a una media teórica. El procedimiento es el siguiente:

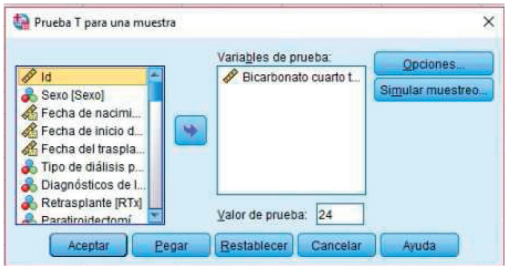
Primero vamos a seleccionar sólo a los pacientes con IRC para explicar este ejemplo (seleccionar casos si IRC=1).

En primer lugar comprobaremos si sigue distribución Normal.

Posteriormente:

Analizar → Comparar medias → Prueba T para una muestra.

Se nos abrirá el siguiente cuadro de diálogo:



Seleccionamos la variable a comparar (en este caso Bicarbonato cuarto trimestre). En "valor de prueba" ponemos el valor descrito en la literatura (en este caso 24). Tras ejecutar la sintaxis, obtenemos:

Estadísticas de muestra única

	N	Media	Desviación estándar	Media de error estándar
HCO34	34	23,997	3,2040	,5495

Prueba de muestra única

	Valor de prueba = 24					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
HCO34	-,005	33	,996	-,0029	-1,121	1,115

En la primera tabla tenemos el valor del bicarbonato cuarto trimestre en nuestros pacientes con IRC (23.997). Vemos que la prueba de significación nos da un valor p=0.996; por tanto, los valores de bicarbonato de nuestros pacientes con IRC no difiere del de los pacientes sin IRC descritos en la literatura.

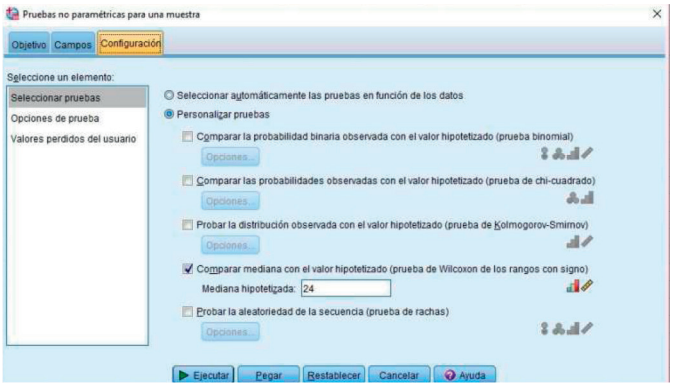
Si no sigue distribución Normal, el procedimiento es diferente. El procedimiento es:

Analizar → Pruebas no paramétricas → Una muestra. La sintaxis es la siguiente:

NPTESTS
/ONESAMPLE TEST (HCO34) WILCOXON(TESTVALUE=24)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.

Se nos abrirán los siguientes cuadros de diálogo:

En la pestaña "Objetivo" seleccionamos "personalizar análisis". En la pestaña "campos" debemos tener cuidado. Vemos que en la primera imagen en el campo "Campos de pruebas" aparecen todas las variables. Debemos seleccionar todas las variables excepto la que nos interesa y pasarlas al campo de la izquierda con la flecha que hay en medio. En la pestaña "configuración" seleccionamos la opción "Comparar mediana con la prueba de Wilcoxon y escribimos el valor de la mediana observada en la literatura (en este caso hemos escrito de nuevo el valor de la media, que si en la población descrita en la literatura sigue unas distribución Normal, el valor de la media y de la mediana coinciden).



Tras ejecutar la sintaxis el resultado es el siguiente:

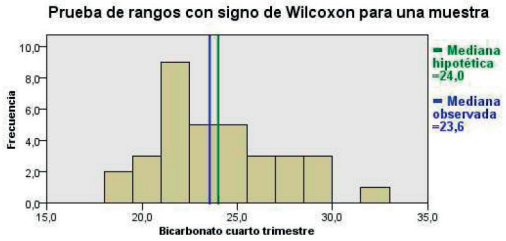
Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La mediana de Bicarbonato cuarto trimestre es igual a 24,0.	Prueba de rangos con signo de Wilcoxon para una muestra	,824	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

El valor es no significativo (p=0.824) y directamente nos indican que conservemos la hipótesis nula, es decir, que la mediana de bicarbonato cuarto trimestre es igual a 24.0.

Si hacemos doble Clic sobre esta tabla, se nos abrirá una nueva ventana con el resumen del modelo y un gráfico ilustrativo donde señalan la mediana observada y la teórica.



N total	34
Estadístico de prueba	284,500
Error estándar	58,484
Estadístico de prueba estandarizado	-,222
Significación asintótica (prueba bilateral)	,824

En algunas versiones es posible que no funcione este procedimiento. En este caso debemos utilizar el procedimiento:

Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos → 2 muestras relacionadas.

Previo a ello debemos crear una variable que contenga el valor observado en la literatura. Vamos a crearla con el procedimiento Compute:

COMPUTE Bicarbo24=24.

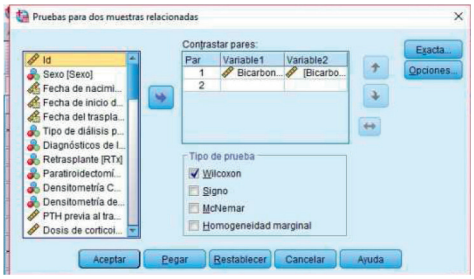
EXECUTE.

Tras ejecutar lo anterior se nos habrá creado en la tabla de datos una variable nueva con el valor 24 en todos los pacientes.

Ahora realizamos el paso descrito:

Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos → 2 muestras relacionadas.

Se nos abrirá el siguiente cuadro de diálogo, donde desplazaremos hacia los campos de la derecha las 2 variables a comparar, con la flecha del medio, y marcamos la opción "Wilcoxon".



La sintaxis es:

NPAR TESTS
/WILCOXON=HCO34 WITH Bicarbo24 (PAIRED)
/MISSING ANALYSIS.

Tras ejecutar la sintaxis, el resultado es el siguiente:

Estadísticos de prueba[a]

	Bicarbo24 - HCO34
Z	-,222[b]
Sig. asintótica (bilateral)	,824

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos negativos.

Vemos que el resultado es el mismo que previamente y por tanto podemos concluir que los valores de bicarbonato de nuestros pacientes en la muestra observada presenta valores similares a los que ya hay descritos en la literatura.

COMPARACIÓN DE UNA VARIABLE CUANTITATIVA EN MÁS DE 2 GRUPOS DISTINTOS.

En el apartado anterior hemos comparado los valores de una variable cuantitativa entre 2 grupos (con y sin IRC), pero en ocasiones vamos a querer comparar los valores de esta variable en más de 2 grupos. Por ejemplo, queremos ver si los valores de bicarbonato al año son diferentes según el estadio de IRC (en nuestra muestra tenemos pacientes: sin IRC, estadio 3, estadio 4 y estadio 5). Se nos podría ocurrir ir seleccionando sólo los estadios a comparar (por ejemplo seleccionar estadio 3 y estadio 4 para compararlos obviando los demás), pero esta no es la manera correcta de hacerlo.

En esta situación no se utiliza la comparación de medias con un t-Student sino que se realiza mediante un análisis de la varianza denominado ANOVA que sigue una ley de Snedecor (F de Snedecor). Vamos a ver cómos se realiza esta prueba con SPSS; los pasos son los siguientes:

Analizar → Comparar medias → Anova de un factor. Se nos abrirá un cuadro de diálogo como el siguiente:

En el cuadro de "Lista de dependientes" seleccionamos las variables a comparar, en este caso "bicarbonato en el cuarto trimestre". En el cuadro "Factor" seleccionamos la variable en cuyos grupos queremos hacer la comparación, en este caso "Estadios IRC". En el botón de la derecha "Opciones" podemos seleccionar que nos muestre además los estadísticos

descriptivos y que nos represente gráficamente las medias (aunque no es el mejor gráfico para ello puesto que sería mejor un diagrama de cajas). La sintaxis es la siguiente:

ONEWAY HCO34 BY EstadIRC
/STATISTICS DESCRIPTIVES
/PLOT MEANS
/MISSING ANALYSIS.

Y los resultados son los siguientes:

Descriptivos

HCO34

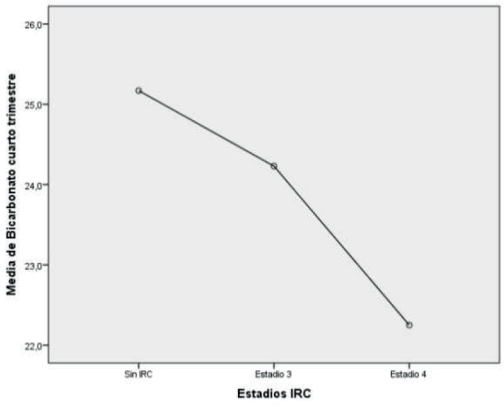
	N	Media	Desviación estándar	Error estándar	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
Sin IRC	27	25,170	2,4542	,4723	24,200	26,141	20,3	30,8
Estadio 3	30	24,230	3,0519	,5572	23,090	25,370	19,4	32,6
Estadio 4	4	22,250	4,2751	2,1376	15,447	29,053	19,3	28,5
Total	61	24,516	2,9328	,3755	23,765	25,268	19,3	32,6

ANOVA

HCO34

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	34,554	2	17,277	2,081	,134
Dentro de grupos	481,529	58	8,302		
Total	516,084	60			

En la primera tabla vemos la descripción de los valores de bicarbonato entre los distintos estadios de IRC. En la segunda tabla vemos la significación del modelo: F de Snedecor 2,081 con una p=0.134. Es decir, no es significativa, los valores de bicarbonato no difieren entre los grupos en la muestra global. Pero vemos que los valores de bicarbonato para los pacientes "Sin IRC" y los que tienen "Estadio 4" parecen ser diferentes, y si observamos el gráfico de representación de las medias parece existir un descenso de los valores de bicarbonato a medida que aumenta el estadio de IRC:



En este caso a lo mejor nos interesa no sólo ver si los valores son o no diferentes en función del estadio de IRC, sino ver también si existe algún tipo de tendencia, es decir, si los valores de las medias van aumentando o disminuyendo a medida que cambia el estadio de IRC. Esto sólo tiene sentido si la variable independiente presenta categorías ordenadas. Podemos valorar si la tendencia es lineal (creciente o decreciente), cuadrática, cúbica, etc., pero habitualmente sólo se suele estudiar si presentan tendencia lineal. Para realizarlo con SPSS, en el cuadro de diálogo anterior, hacemos Clic en botón "Contrastes" y se nos abrirá el siguiente cuadro de diálogo:

Seleccionamos la opción "Polinómica" y marcamos la opción de la derecha "Grado" "Lineal".

El resultado es el siguiente:

ANOVA

HCO34

			Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	(Combinado)		34,554	2	17,277	2,081	,134
	Término lineal	Ponderados	27,735	1	27,735	3,341	,073
		Desviación	6,820	1	6,820	,821	,369
Dentro de grupos			481,529	58	8,302		
Total			516,084	60			

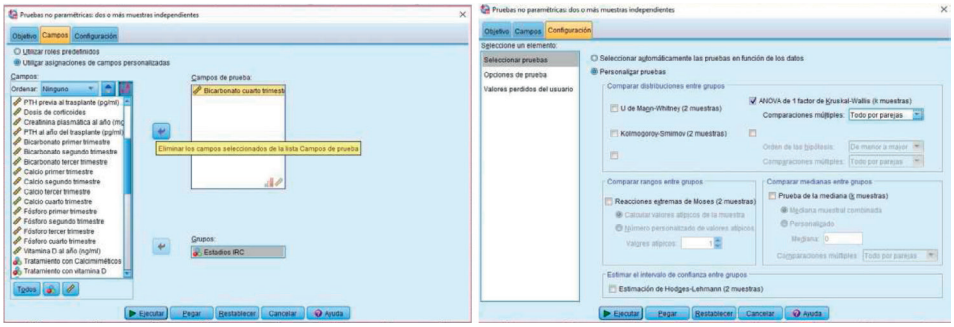
Vemos que la tabla es igual a la anterior pero se le han añadido algunos elementos. Vemos que en el apartado "Entre grupos" aparece el término "Término lineal" con un valor de F de Snedecor de 3,341 y un valor p=0.073. Es decir, existe una tendencia lineal en las medias de bicarbonato en los distintos estadios de IRC pero no llega a ser estadísticamente significativo.

Los términos "Suma de cuadrados", "Media cuadrática" etc., los veremos en los capítulos de regresión de forma más detallada.

Vemos que para el análisis estadístico anterior no hemos hablado sobre la necesidad de que los valores se distribuyan de manera Normal. La prueba de análisis de la varianza es tan potente que no es necesario que los valores se distribuyan de manera Normal, aunque siempre es recomendable comprobar en muestras menores de 30 casos, si la variable Y

(bicarbonato) se distribuye de manera Normal en los distintos subgrupos. En nuestro ejemplo (tras comprobarlo), los datos se distribuyen de manera Normal. En caso de no hacerlo en algún subgrupo, el análisis se realiza mediante el método de Kruskal- Wallis, que además sirve para casos en los que tengamos una tabla de contingencia con una variable de exposición categórica y una variable respuesta con categorías ordenadas. El procedimiento SPSS es el siguiente:

Analizar → Pruebas no paramétricas → Muestras independientes. Se abrirán los siguientes cuadros de diálogos:



En la pestaña "Campos" seleccionamos los respectivos campos, y en la pestaña "Configuración" seleccionamos la opción "ANOVA de 1 factor de Kruskal-Wallis". La sintaxis y resultados son los siguientes:

NPTESTS

/INDEPENDENT TEST (HCO34) GROUP (EstadIRC)

KRUSKAL_WALLIS(COMPARE=PAIRWISE)

/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE

/CRITERIA ALPHA=0.05 CILEVEL=95.

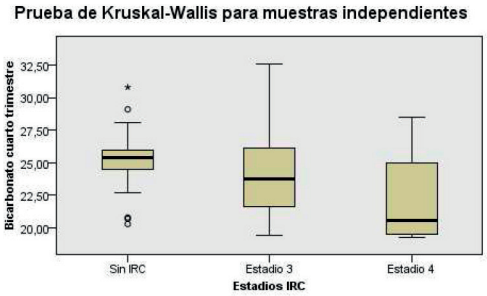
Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de Bicarbonato cuarto trimestre es la misma entre las categorías de Estadios IRC.	Prueba de Kruskal-Wallis para muestras independientes	,130	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

Obtenemos una prueba no significativa con p=0.130, es decir, "La distribución de Bicarbonato al año es la misma entre las categorías de Estadios de IRC".

Si hacemos doble Clic sobre este cuadro se nos abrirá una ventana con información del modelo y un gráfico donde podemos ver de manera visual las diferencias:



N total	61
Estadístico de prueba	4,073
Grados de libertad	2
Significación asintótica (prueba bilateral)	,130

1. Los estadísticos de prueba se ajustan para empates.
2. No se realizan múltiples comparaciones porque la prueba global no muestra diferencias significativas en las muestras.

Si el procedimiento anterior no funciona, tenemos la alternativa de realizarlo con cuadros de diálogos antiguos:

Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos → K muestras independientes, donde se nos abrirá el siguiente cuadro de diálogo:

En este cuadro además tenemos la opción de marcar el procedimiento de Jonckheere-Terpstra, que es necesario para valorar si existe o no tendencia lineal. La sintaxis y resultados son los siguientes:

NPAR TESTS

/K-W=HCO34 BY EstadIRC(1 5)

/J-T=HCO34 BY EstadIRC(1 5)

/MISSING ANALYSIS.

Rangos

	EstadIRC	N	Rango promedio
HCO34	Sin IRC	27	35,28
	Estadio 3	30	28,85
	Estadio 4	4	18,25
	Total	61	

Estadísticos de prueba[a,b]

	HCO34
Chi-cuadrado	4,073
gl	2
Sig. asintótica	,130

a. Prueba de Kruskal Wallis

b. Variable de agrupación: EstadIRC

Prueba de Jonckheere-Terpstra[a]

	HCO34
Número de niveles en EstadIRC	3
N	61
Estadístico J-T observado	378,500
Estadístico J-T de media	519,000
Desviación estándar del estadístico J-T	71,346
Estadístico J-T estándar	-1,969
Sig. asintótica (bilateral)	,049

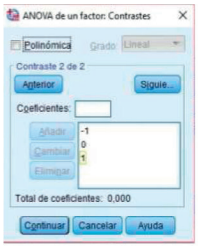
a. Variable de agrupación: EstadIRC

Vemos que el valor de la prueba de Kruskal-Wallis presenta el mismo valor p=0.130. En esta ocasión al valorar la existencia de tendencia lineal mediante la prueba de Jonckheere-Terpstra, sí se obtiene un valor significativo con una p=0.049. Es significativo por muy poco; el valor obtenido mediante la F de Snedecor era de 0.073, pero como se ha comentado anteriormente, en el caso de este ejemplo, la distribución del bicarbonato entre las distintas categorías es Normal, y además hemos dicho que la prueba de ANOVA es muy potente y no se necesita que presenten distribución Normal, aunque nunca está demás el comprobarlo y obtener resultados de manera "elegante".

Hasta ahora hemos comprobado si los valores de la variable dependiente se distribuyen de manera distinta entre las distintas categorías (de manera global) y si presentan además cierta tendencia. Pero es posible que también nos interese comprobar entre qué categorías se encuentran estas diferencias. Para ello vamos a utilizar los "Contrastes". Los contrastes asignan valores (coeficientes) a las distintas categorías para poder hacer comparaciones entre ellas. En general, y sobre todo si las categorías son ordenadas, se suelen comparar con una categoría de referencia. En nuestro ejemplo consistiría en comparar los valores de bicarbonato de los estadios más avanzados con los del estadio "Sin IRC" que sería tomado como de referencia. Pero se pueden comparar categorías entre sí o grupos de categorías (por ejemplo el estadio "Sin IRC" y "estadio 3" compararlos en conjunto con los "estadio 4" y "estadio 5"). ¿Qué valores hay que darles a los coeficientes de cada categoría? La categoría que va a ser la de referencia va a tener el valor (-1) y la categoría a comparar el valor (1); al resto de categorías se les asigna el valor 0 (para que no se tengan en cuenta en la comparación). Por ejemplo, si queremos comparar la categoría "estadio 3" con la categoría "Sin IRC", la categoría "estadio 3" tendría el valor 1 y la categoría "Sin IRC" el valor -1. A las categorías "estadio 4" y "estadio 5" se les daría el valor 0 para que no "entren en el juego de las comparaciones". Si queremos comparar las categorías "Sin IRC" y "estadio 3" con las categorías "estadio 4" y "estadio 5", a las 2 primeras les daríamos los valores -1/2 y -1/2, y a las 2 últimas los valores ½ y ½. Si vamos a comparar el "estadio Sin IRC" con los "estadios 4" y "estadio 5", al primero le damos el valor -1, al "estadio 4" le damos el valor ½ y al "estadio 5" también 1/2, mientras que al "estadio 3" le damos el valor 0. El procedimiento con SPSS es el similar a los pasos previos de ANOVA pero al hacer Clic en el botón "Contrastes" debemos asignar el valor de los coeficientes. Vamos a comparar cada estadio con el estado "Sin IRC" que vamos a tomar como referencia. La sintaxis es la siguiente:

```
ONEWAY HCO34 BY EstadIRC
/CONTRAST=-1 1 0
/CONTRAST=-1 0 1
/STATISTICS HOMOGENEITY
/MISSING ANALYSIS.
```

El cuadro de contrastes resultantes y los resultados son los siguientes:



Coeficientes de contraste

Contraste	EstadIRC		
	Sin IRC	Estadio 3	Estadio 4
1	-1	1	0
2	-1	0	1

Prueba de homogeneidad de varianzas

HCO34

Estadístico de Levene	gl1	gl2	Sig.
1,707	2	58	,190

Pruebas de contraste

		Contraste	Valor de contraste	Error estándar	t	gl	Sig. (bilateral)
HCO34	Suponer varianzas iguales	1	-,940	,7644	-1,230	58	,224
		2	-2,920	1,5437	-1,892	58	,064
	No se asume varianzas iguales	1	-,940	,7304	-1,287	54,349	,203
		2	-2,920	2,1891	-1,334	3,299	,267

En nuestro ejemplo, no hay ningún paciente en "estadio 5" por lo que este contraste no se puede realizar, y por eso sólo hay 2 tipos. Vemos en la primera tabla el valor de los coeficientes asignados.

En la segunda tabla, en la primera columna nos informa sobre la diferencia del bicarbonato al comparar cada categoría. Vemos que existen 2 filas de resultados (la que supone varianzas iguales y las que no); vemos que en el estadístico de Levene (que valora la igualdad de varianzas), la significación es p=0.190 (no significativo) por lo que asumimos que las varianzas son iguales y por tanto tenemos en cuenta los datos de la primera fila. Asumiendo varianzas iguales vemos que al comparar el "estadio 4" con el estadio "Sin IRC" (contraste 2) se obtiene una p casi significativa 0.064, mientras que el contraste 1 es claramente no significativo (el que compara el "estadio 3" con el estadio "sin IRC"). La diferencia entre el "estadio 4" y el estadio "sin IRC" (contraste 2) es de 2.920, es decir, al restar los valores de bicarbonato entre el "estadio 4" y el estadio "sin IRC" es de casi 3 mEq/l.

Al igual que en anteriores análisis estadísticos, la significación global de la prueba para comparar más de 2 medias se establece en un valor p=0.05. Sin embargo, en este tipo de análisis existe una peculiaridad. Está demostrado que a medida que aumenta el número de comparaciones (es decir de contrastes), la probabilidad de que alguna comparación sea significativa ($p<0.05$) aumenta a medida que aumenta el número de contrastes. Dicho de otra manera, si sumásemos los valores "p" de cada contraste, el valor final sería mayor de 0.05. Por ello, para que el valor global de la prueba sea de una $p<0.05$ se necesita que el valor de significación de cada contraste sea menor a 0.05, y a

medida que aumenta el número de comparaciones, menor tiene que ser el valor "p" para considerarlo como estadísticamente significativo. Para ello debemos realizar correcciones, llamada *Corrección de Bonferroni*.

En la siguiente tabla se muestra el valor que debe tener cada comparación para considerarla significativa según el número de comparaciones:

Número de comparaciones	Valor "p" requerido para una "p" global <0.05
1	0.05
2	0.0253
3	0.0170
4	0.0127
5	0.0102

Si realizamos 4 comparaciones, el valor "p" de cada comparación debe ser menor de 0.0127 para considerarlo como estadísticamente significativo, pues así nos aseguramos que el valor global del modelo sea menor a 0.05.

Para realizar con SPSS la corrección de Bonferroni debemos marcar la opción en el cuadro de diálogo del análisis de la varianza (ANOVA de un factor) en el botón "Post Hoc" y seleccionar la opción "Bonferroni". La tabla resultante es la siguiente:

Comparaciones múltiples

Variable dependiente: HCO34

Bonferroni

(I) EstadIRC	(J) EstadIRC	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Sin IRC	Estadio 3	,9404	,7644	,671	-,944	2,825
	Estadio 4	2,9204	1,5437	,191	-,885	6,726
Estadio 3	Sin IRC	-,9404	,7644	,671	-2,825	,944
	Estadio 4	1,9800	1,5337	,605	-1,801	5,761
Estadio 4	Sin IRC	-2,9204	1,5437	,191	-6,726	,885
	Estadio 3	-1,9800	1,5337	,605	-5,761	1,801

Vemos que en la tabla aparecen todos los contrastes posibles, con signo positivo y negativo, por lo que se repiten. En la primera fila aparecen los valores que habíamos obtenido previamente, pero vemos que cambia el valor de la significación. En este caso no hay ningún contraste significativo. De haber alguno estaría señalado con un asterisco (*) y debajo de la tabla aparecería una leyenda indicando que el nivel de significación se ha obtenido para p<0.05 (que sería la significación global del modelo).

ANÁLISIS ESTADÍSTICO BÁSICO CON VARIABLES CATEGÓRICAS

COMPARACIÓN DE VARIABLES CATEGÓRICAS CON 2 CATEGORÍAS

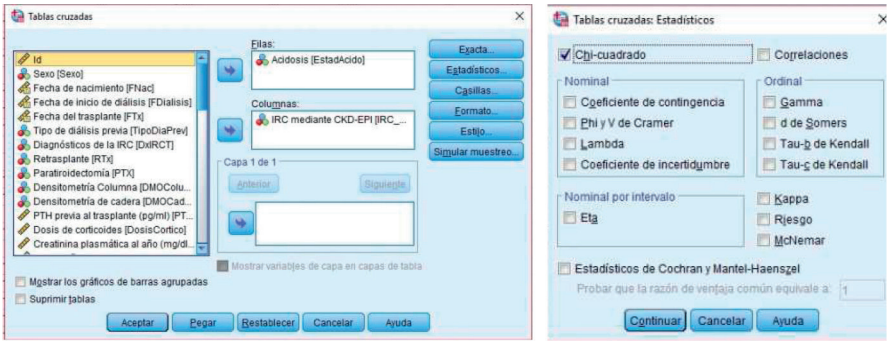
En este apartado vamos a ver los procedimientos estadísticos necesarios para comparar variables categóricas. Vamos a partir del modelo más sencillo donde se comparan 2 variables categóricas ambas binarias formando una tabla de 2 x 2. El ejemplo en nuestra base de datos sería: ¿la proporción de pacientes con acidosis metabólica es diferente según los pacientes tengan o no IRC? La tabla resultante 2 x 2 de este ejemplo sería la siguiente:

Acidosis	Presencia de IRC		Total
	No	Sí	
No	23	15	38
Sí	4	19	23
Total	27	34	61

Para la elaboración de la anterior tabla debemos colocar la variable exposición en columnas (Presencia de IRC) y la variable respuesta (Acidosis) en filas, similar a un eje de coordenadas, donde la abscisa (eje X) es la variable exposición y la ordenada (eje Y) la variable respuesta, poniendo las columnas en orden creciente.

Vemos que esta tabla así no nos dice nada. Podemos intuir que el número de pacientes con acidosis en el grupo con IRC es mayor que en el grupo sin IRC (19 frente a 4), pero esto no basta para dar conclusiones. Para comprobar si estas diferencias son significativas se utiliza es estadístico Chi-Cuadrado (χ^2). Este estadístico compara si la proporción de pacientes con acidosis en el grupo con IRC es diferente al de los pacientes del grupo sin IRC. Para ello parte del cálculo de cuáles serían las frecuencias esperadas en cada casilla en caso de no existir diferencias entre ambos grupos y comparar las frecuencias esperadas con las observadas a través del estadístico de contraste χ^2. Veamos la siguiente tabla elaborada con SPSS sobre el ejemplo anterior. El procedimiento sería el siguiente:

Analizar → Estadísticos Descriptivos → Tablas cruzadas. Y nos saldría el siguiente cuadro de diálogo:



En el primer cuadro de diálogo seleccionamos la variable exposición y la desplazamos al cuadro "Columnas" y la variable respuesta al cuadro "Filas" tal como hemos explicado anteriormente. En el botón de la derecha "Estadísticos" seleccionamos la opción "Chi- cuadrado". Si hacemos Clic en el botón "Casillas" obtenemos el siguiente cuadro de diálogo, donde indicaremos que nos aparezcan las frecuencias (Recuentos) observadas y esperadas, así como los porcentajes en columnas. La sintaxis es la siguiente:

```
CROSSTABS
/TABLES=EstadAcido BY IRC_CKD
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT EXPECTED COLUMN
/COUNT ROUND CELL.
```

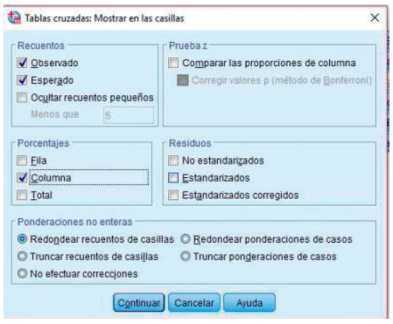


Tabla cruzada EstadAcido*IRC_CKD

			IRC_CKD		
			No	Sí	Total
EstadAcido	No	Recuento	23	15	38
		Recuento esperado	16,8	21,2	38,0
		% dentro de IRC_CKD	85,2%	44,1%	62,3%
	Sí	Recuento	4	19	23
		Recuento esperado	10,2	12,8	23,0
		% dentro de IRC_CKD	14,8%	55,9%	37,7%
Total		Recuento	27	34	61
		Recuento esperado	27,0	34,0	61,0
		% dentro de IRC_CKD	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	10,806[a]	1	,001		
Corrección de continuidad[b]	9,128	1	,003		
Razón de verosimilitud	11,523	1	,001		
Prueba exacta de Fisher				,001	,001
Asociación lineal por lineal	10,629	1	,001		
N de casos válidos	61				

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 10,18.

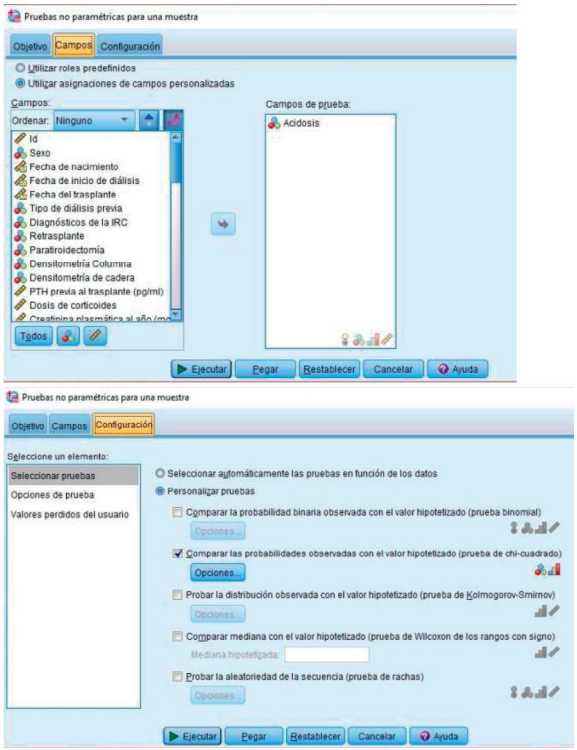
b. Sólo se ha calculado para una tabla 2x2

Vemos que esta tabla es similar a la que habíamos elaborado anteriormente, donde aparecen además en cada casilla las frecuencias esperadas. En la primera casilla (pacientes sin acidosis y sin IRC) tenemos 23 pacientes, pero la frecuencia esperada si no hubiese diferencias, sería de 16.8 pacientes (es decir, hemos obtenido más pacientes de los que serían esperados). Lo mismo sucede en la última casilla (pacientes con Acidosis y con IRC) donde hemos obtenido 19 pacientes y sin embargo esperaríamos encontrar tan sólo 12.8 pacientes. En cada casilla además aparecen los porcentajes observados. En los pacientes con IRC existe un 55.9% de pacientes con Acidosis, mientras que en los pacientes sin IRC es tan sólo del 14.8%. En la segunda tabla obtenemos la significación de esta diferencia; vemos que es estadísticamente significativo pues el valor de Chi-cuadrado 10,806 es significativo con una p=0.001.

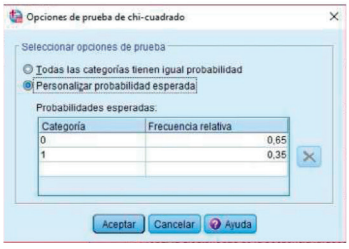
Vemos que en esta segunda tabla aparecen otros valores de significación. Sólo prestaremos atención de momento al valor de la "Prueba exacta de Fisher". Para que la prueba de Chi-cuadrado pueda realizarse, el número mínimo de casos esperados en cada casilla debe ser al menos de 5. Si en alguna casilla hay menos de 5 recuentos esperados debemos utilizar la significación de la "Prueba exacta de Fisher". Esto vendrá reflejado debajo de la tabla. En este caso vemos que hay 0 casillas (0.0%) con recuento esperado menor de 5 y por tanto la prueba de Chi-cuadrado es válida.

En determinados estudios no vamos a querer comparar las diferencias de frecuencias entre 2 grupos, sino que vamos a querer ver si la frecuencia que nosotros hemos observado en nuestro estudio es diferente de otra frecuencia que se haya publicado en otros trabajos. Por ejemplo, en nuestro estudio queremos ver si la frecuencia de acidosis metabólica observada es diferente de la ya publicada en otros estudios, y hemos visto en publicaciones que se describen porcentajes de acidosis del 35%. El procedimiento con SPSS sería el siguiente:

Analizar → Pruebas no paramétricas → Una muestra. Se nos abrirá la siguiente ventana:



En la pestaña "Campos", al igual que sucedía cuando queríamos comparar una mediana observada a una teórica, debemos pasar todas las variables, excepto la que nos interesa, hacia el cuadro de la izquierda. En la pestaña "Configuración" seleccionamos la opción "Comparar las probabilidades observadas con el valor hipotetizado (prueba de chi- cuadrado)". En el botón opciones ponemos el valor de las categorías (nosotros las categorías "Sin acidosis" le dimos el valor 0, y "Con acidosis" el valor 1. En los campos "Frecuencia relativa" ponemos el valor de las probabilidades observadas en los estudios que hayamos leído (en este caso hemos dicho que está publicada una probabilidad del 35% de acidosis, por tanto a la categoría 1 le daremos el valor 0.35. El cuadro es el siguiente y tras ejecutar la sintaxis obtenemos el siguiente cuadro de resultados:



Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	Las categorías de Acidosis se producen con probabilidades especificadas.	Prueba de chi-cuadrado para una muestra	,658	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

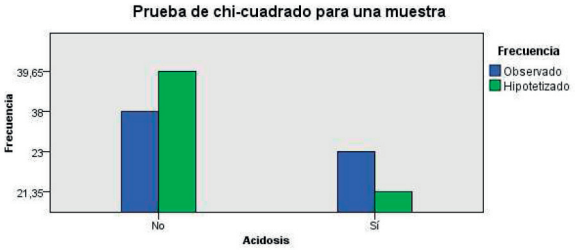
Sintaxis completa:

NPTESTS
/ONESAMPLE TEST (EstadAcido)

CHISQUARE(EXPECTED=CUSTOM(CATEGORIES=0 1 FREQUENCIES=0.65 0.35))
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.

Vemos que el valor obtenido no es significativo con una p=0.658, es decir, nuestra frecuencia observada es la misma que la descrita en las publicaciones.

Si hacemos doble Clic sobre el cuadro de resultados, obtenemos información resumen del modelo y un gráfico de barras donde podemos ver de manera visual nuestra probabilidad observada y la teórica:



N total	61
Estadístico de prueba	,196
Grados de libertad	1
Significación asintótica (prueba bilateral)	,658

1. Hay 0 casillas (0%) con valores esperados menores que 5. El valor esperado mínimo es 21,350.

Si el procedimiento anterior no funciona, tenemos la alternativa de realizarlo con cuadros de diálogos antiguos:

Analizar → Pruebas no paramétricas → Cuadro de diálogos antiguos → Chi-Cuadrado, donde se nos abrirá el siguiente cuadro de diálogo:

El cuadro "Lista de variables de prueba" seleccionamos la variable a comparar. En el cuadro "Valores esperados" debemos ir introduciendo los valores que hemos encontrado publicados haciendo Clic sobre el botón "Añadir". Debemos tener cuidado e ir introduciendo los valores en orden ascendente de categorías, es decir, en primer lugar la categoría 0 y luego la categoría 1, y así sucesivamente en caso de haber más categorías.

En este caso como hemos visto en los estudios que la frecuencia de acidosis es del 35%, la categoría sin acidosis tendrá el valor 0.65 que introduciremos en primer lugar (pues es la categoría 0) y posteriormente el valor 0.35. La sintaxis y el resultado son los siguientes:

NPAR TESTS
/CHISQUARE=EstadAcido
/EXPECTED=0.65 0.35
/MISSING ANALYSIS.

Estadísticos de prueba

	EstadAcido
Chi-cuadrado	,196[a]
gl	1
Sig. asintótica	,658

a. 0 casillas (0,0%) han esperado frecuencias menores que 5. La frecuencia mínima de casilla esperada es 21,3.

Vemos que el resultado es igual al obtenido con el otro procedimiento.

COMPARACIÓN DE VARIABLES CATEGÓRICAS CON MÁS DE 2 CATEGORÍAS

Para la comparación de una variable categórica binaria entre varias categorías, el procedimiento es similar al de una tabla 2 x 2 utilizando igualmente Chi-cuadrado o estadístico exacto de Fisher según corresponda. Por ejemplo, queremos comparar si la proporción de acidosis metabólica es distinta según los estadios de IRC. El procedimiento con SPSS es similar al de la tabla 2 x 2:

Analizar → Estadísticos Descriptivos → Tablas cruzadas. La sintaxis correspondiente sería:

CROSSTABS
/TABLES=EstadAcido BY EstadIRC
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT EXPECTED COLUMN
/COUNT ROUND CELL.

Tras ejecutar la sintaxis obtenemos la siguiente tabla:

Tabla cruzada EstadAcido*EstadIRC

			EstadIRC			
			Sin IRC	Estadio 3	Estadio 4	Total
EstadAcido	No	Recuento	23	14	1	38
		Recuento esperado	16,8	18,7	2,5	38,0
		% dentro de EstadIRC	85,2%	46,7%	25,0%	62,3%
	Sí	Recuento	4	16	3	23
		Recuento esperado	10,2	11,3	1,5	23,0
		% dentro de EstadIRC	14,8%	53,3%	75,0%	37,7%
Total		Recuento	27	30	4	61
		Recuento esperado	27,0	30,0	4,0	61,0
		% dentro de EstadIRC	100,0%	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	11,511[a]	2	,003
Razón de verosimilitud	12,231	2	,002
Asociación lineal por lineal	11,315	1	,001
N de casos válidos	61		

a. 2 casillas (33,3%) han esperado un recuento menor que 5. El recuento mínimo esperado es 1,51.

En la primera tabla vemos los recuentos esperados si la probabilidad de acidosis metabólica fuese la misma en las tres categorías de IRC. Vemos que los porcentajes de pacientes con acidosis metabólica son diferentes según el estadio de IRC siendo estadísticamente significativo según la prueba Chi-cuadrado con un valor p=0.003. Vemos debajo de la segunda tabla que existen 2 casillas con un recuento esperado inferior a 5 y por tanto teóricamente la prueba de Chi-cuadrado no podría aplicarse; sin embargo cuando existen más de 2 categorías ordenadas esta condición no es tan necesaria exigiéndose por lo general que el número de casillas con recuento esperado inferior a 2 no sea elevado y que las casillas con este recuento inferior a 2 no sean contiguas. En este caso sólo hay una casilla con un recuento esperado inferior a 2 (recuento esperado 1.5). En caso de haber muchas casillas con recuento inferior a 2 debemos replantearnos si la comparación que queremos hacer es factible con el tamaño de nuestra muestra. De ser factible, podemos usar la significación de la Razón de verosimilitud, cuyo valor es muy parecido al de Chi-cuadrado. Pero generalmente, cuando existen muchas casillas con recuentos esperados inferiores a 2 es porque nuestra muestra es pequeña para realizar esa comparación; una solución a este problema en caso de no poder aumentar el tamaño de muestra es agrupar categorías.

En el caso del ejemplo anterior no sólo nos va a interesar saber que el porcentaje de acidosis cambia según el estadio de IRC, sino que además nos puede interesar saber si existe una tendencia. Esto sólo tiene sentido para categorías ordenadas. La pregunta a contestar es "si a medida que empeora el estadio de IRC también se modifica el porcentaje de acidosis". En la tabla primera vemos que los porcentajes van aumentando a medida que aumenta el estadio de IRC, pero ¿esta tendencia es significativa? Para contestar esta pregunta nos debemos fijar en la segunda tabla en la línea "Asociación lineal por lineal" vemos que su valor es significativo con un valor p=0.001, es decir, existe una tendencia lineal. Por tanto en el anterior ejemplo podemos decir que el porcentaje de acidosis se modifica con los estadios de IRC y que además existe una tendencia lineal con un mayor porcentaje de acidosis a medida que empeora el estadio de IRC. La tendencia positiva o negativa debemos observarla en cómo se modifiquen los porcentajes o bien a través de un gráfico.

Hasta ahora hemos hecho comparaciones entre una variable dependiente con 2 categorías (Acidosis Sí/No) con variables independientes con 2 categorías (IRC Sí/No) o más de 2 categorías (Estadios de IRC). Cuando la comparación se realiza con variables dependientes con más de 2 categorías, el problema se complica demasiado. Imaginemos que queremos comparar los estadios de IRC según el diagnóstico (es sólo un ejemplo, pues la comparación no tiene sentido). En este caso lo que se suele hacer es realizar subtablas 2 x 2 con las distintas

categorías de la variable dependiente e independiente creadas con procedimientos COMPUTE y ver en cuál o cuáles de ellas existen las diferencias.

Al igual que sucedía al comparar varias medias, si la variable independiente es categórica pero no ordenada (es decir no tiene sentido ver la tendencia lineal), podemos querer ver entre qué categorías existen las diferencias. En el ejemplo anterior (que no sería aplicable porque son categorías ordenadas pero vamos usarlo como ejemplo) podríamos querer ver entre qué estadio de IRC existen las diferencias. Al igual que con la comparación de medias debemos hacer la corrección de Bonferroni, usando valores "p" en cada comparación inferiores para que el valor "p" de la comparación global sea inferior a 0.05. En la siguiente tabla aparecen los valores "p" según el número de comparaciones:

Número de comparaciones	Valor "p" requerido para una "p" global <0.05
1	0.05
2	0.0253
3	0.0170
4	0.0127
5	0.0102

MEDIDAS DE ASOCIACIÓN ENTRE VARIABLES CATEGÓRICAS

En el apartado anterior hemos comprobado si existen diferencias significativas entre variables categóricas, e incluso hemos valorado la posibilidad de que exista cierta tendencia a medida que cambia las categorías de la variable independiente. Pero probablemente esto no nos dé toda la información que nos gustaría. No nos contesta a la pregunta por ejemplo "¿cuántas veces es más frecuente la acidosis metabólica entre los pacientes con IRC respecto de los pacientes sin IRC?" Esta pregunta se contesta mediante medidas de asociación, las cuáles se van a realizar generalmente mediante análisis de Regresión que veremos más adelante, pero también es posible realizarlas mediante tablas de contingencia.

Las medidas de asociación van a depender del diseño del estudio que hayamos realizado. Si lo que hemos hecho no ha sido un Ensayo Clínico, las medidas de asociación van a ser medidas "crudas" entre las variables a comparar, pero se pueden ver modificadas por la presencia de otras variables. La capacidad de dar una medida de asociación "neta" una vez extraído el efecto de otras variables sólo va a ser posible hacerlo mediante Regresión.

Vamos a tomar como ejemplo el análisis de la asociación entre presentar o no acidosis metabólica con la presencia de Osteoporosis. Esta variable aún no está calculada en la Tabla de datos y por tanto debemos calcularla antes de continuar. ¿Cómo se realizaría? Vamos a considerar "Paciente Osteoporótico" aquél que tenga osteoporosis bien en columna o bien en cadera. La sintaxis exhaustiva sería:

IF (DMOColumna=2 OR DMOCadera=2) Osteop=1.

IF (DMOColumna=0 AND DMOCadera=0) Osteop=0.

IF (DMOColumna=1 AND DMOCadera=0) Osteop=0.

```
IF (DMOColumna=1 AND DMOCadera=1) Osteop=0.
IF (DMOColumna=0 AND DMOCadera=1) Osteop=0.
IF (DMOColumna=$sysmis AND DMOCadera<2) Osteop=$sysmis.
IF (DMOColumna<2 AND DMOCadera=$sysmis) Osteop=$sysmis.
EXECUTE.
```

Después del procedimiento anterior siempre interesa realizar un "LIST VARIABLES" con las variables en cuestión para comprobar que está todo correcto. Por supuesto después debemos definir las variables.

Una vez realizado lo anterior vamos a crear la tabla 2 x 2 resultante de "Osteoporosis" y "Acidosis". La variable exposición es acidosis y respuesta osteoporosis.

Tabla cruzada Osteop*EstadAcido

			EstadAcido		Total
			No	Sí	
Osteop	No	Recuento	33	15	48
		% dentro de EstadAcido	89,2%	65,2%	80,0%
	Sí	Recuento	4	8	12
		% dentro de EstadAcido	10,8%	34,8%	20,0%
Total		Recuento	37	23	60
		% dentro de EstadAcido	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	5,094[a]	1	,024		
Corrección de continuidad[b]	3,706	1	,054		
Razón de verosimilitud	4,980	1	,026		
Prueba exacta de Fisher				,044	,028
Asociación lineal por lineal	5,009	1	,025		
N de casos válidos	60				

a. 1 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 4,60.

b. Sólo se ha calculado para una tabla 2x2

Vemos que los porcentajes de osteoporosis son diferentes según haya o no acidosis metabólica, y que esta diferencia es estadísticamente significativa con una p=0.044 según la Prueba exacta de Fisher.

Sobre esta tabla se pueden realizar distintos cálculos:

Osteoporosis	Presencia de Acidosis		Total
	No	Sí	
No	33	15	48
Sí	4	8	12
Total	37	23	60

- Proporción de enfermedad entre los expuestos: 8/23=0.35.
- Proporción de enfermedad entre los NO expuestos: 4/37=0.11.

- Odds de enfermedad entre los expuestos: 8/15=0.53.
- Odds de enfermedad entre los NO expuestos: 4/33=0.12.
- Proporción de expuestos entre los enfermos: 8/12=0.67.
- Proporción de expuestos entre los NO enfermos: 15/48=0.31.
- Odds de expuestos entre los enfermos: 8/4= 2.
- Odds de expuestos entre los NO enfermos: 15/33= 0.45.

Según el tipo de estudio podremos hablar de una medida de asociación u otra. Si es un estudio transversal, las proporciones se denominan Prevalencias, representando a todos los enfermos o expuestos existentes en un momento determinado del tiempo, y podremos hablar de cualquier medida de asociación de las anteriores.

Si es un estudio de Cohortes o ensayo clínico, las proporciones se denominan Riesgos o Incidencias acumuladas. Los pacientes se eligen al comienzo del estudio según presenten o no la exposición y los seguiremos en el tiempo para ver si desarrollan la enfermedad. Por ello, al comienzo del estudio los pacientes que ya están enfermos deben ser descartados. Se van a contabilizar los pacientes que desarrollan la enfermedad durante el periodo de estudio y por ello van a ser casos incidentes. El objetivo será comparar el número de enfermos en el grupo de expuesto frente al grupo no expuesto. Los pacientes al comienzo del estudio al no estar enfermos, van a estar en "riesgo" de desarrollar la enfermedad durante el periodo de seguimiento y por ello se denomina a la proporción de enfermos "Riesgo o Incidencia acumulada".

De los cálculos anteriores se pueden extraer diferencias y razones:

- Diferencias de Proporciones (Prevalencias o Riesgos).
- Razón de proporciones (Prevalencias o Riesgos).
- Razón de Odds (Prevalencias o Riesgos).

Estudio Transversal

Son datos que se recogen en un momento determinado del tiempo y se supone que representan tanto a la exposición de la población como a la enfermedad en la población. Podemos hablar de cualquiera de las anteriores medidas de asociación: Diferencias de Prevalencias de enfermedad o de exposición, Razones de enfermedad o exposición y Razones de Odds de enfermedad o exposición.

Estudio Cohortes o Ensayo Clínico

Estos tipos de estudios se eligen a los sujetos según presenten la exposición o no. Sería el ejemplo de nuestra base de datos. Dado que los sujetos se eligen según exposición, no tiene sentido hablar de proporciones u Odds de exposición entre los enfermos. Las medidas de asociación serían:

- Diferencia de Riesgos de enfermedad (DR): 8/23 − 4/37 = 0.24 Los pacientes con Acidosis tienen un 24% más de Osteoporosis.

- Razón de Odds de enfermedad (OR):(8/15)/(4/33) = 4.4 La Odds de osteoporosis es 4.4 veces mayor entre los pacientes con Acidosis que en entre los pacientes sin Acidosis.
- Razón de Riesgo de enfermedad (RR): (8/23)/(4/37) = 3.22 Los pacientes con Acidosis tienen 3.22 veces más riesgo de Osteoporosis que los pacientes sin Acidosis.

Las Odds y las Razones de Odds son difíciles de interpretar aunque su interpretación es similar al de la Razón de Proporciones. En este caso la Razón de Proporción se denomina Razón de Riesgos (RR) o Riesgo Relativos, que será la medida de asociación a utilizar en estos estudios, además de la diferencia de Riesgos. Sin embargo mediante Regresión como veremos más adelante sólo se pueden extraer Odds y Razones de Odds.

La Razón de Odds siempre va a dar valores más alejados de 1 que la Razón de Riesgo. Es decir, siempre va a sobreestimar la asociación. En este ejemplo vemos que la RR es 3.22 y que la OR es 4.4. Si hubiese sido RR: 0.8, la OR hubiese sido menor.

Los valores de OR y RR varían entre 0 e ∞. Una OR o RR de 1 indica que no hay asociación (el valor del numerador es igual que el denominador). Si es mayor de 1 indica que hay aumento del riesgo. Si es menor de 1 indica que disminuye el riesgo, es decir, es un factor protector.

Los valores de la DR puede ser positivo o negativo: si es positivo habrá aumento de la proporción de riesgo y si es negativo habrá una disminución de la proporción de riesgo (factor protector).

Estudio Casos-Controles

Estos tipos de estudios se eligen a los sujetos según presenten la enfermedad o no. Dado que los sujetos se eligen según enfermedad, no tiene sentido hablar de proporciones u Odds de enfermedad entre los expuestos y por tanto no vamos a poder calcular Riesgos de enfermedad en los expuestos. La única medida de asociación que se puede realizar es:

- Razón de Odds de exposición (OR):(8/4)/(15/33) = 4.4 La Odds de exposición es 4.4 veces mayor entre los pacientes con Osteoporosis que en entre los pacientes sin Osteoporosis.

Mediante una tabla de contingencia tan sólo se pueden extraer Riesgos. Dado que nuestro ejemplo es un estudio de Cohortes, vamos a ver cómo se realizaría. En primer lugar, SPSS para poder hacerlo debe modificar la estructura de toda la tabla; debemos poner en columna la respuesta y en filas la exposición. Además, SPSS calcula Riesgos a través de la categoría más baja, lo cual no tiene mucho sentido puesto que el sentido de los estudios es calcular asociaciones al pasar de una categoría inferior a una superior (pasar de 0 a 1). Por tanto, debemos recodificar la variable (a la categoría 0 darle el valor 1 y a la categoría 1 darle el valor 0). Esto se hace con el procedimiento AUTORECODE y creando 2 variables nuevas con las mismas etiquetas que las originales. El procedimiento es:

Transformar → Recodificación automática. Se nos abrirá un cuadro de diálogo como el siguiente:

Seleccionamos las 2 variables a estudio y escribimos el nombre de la nueva variable en el recuadro "Nuevo nombre". Por último seleccionamos la opción "Recodificar empezando desde Mayor valor" y de esta forma los recodifica a la inversa de cómo lo teníamos previamente.

Posteriormente realizamos de nuevo la tabla de contingencia poniendo en columnas la respuesta y en las filas la exposición (a la inversa de cómo hacemos habitualmente). En el botón "Estadístico" seleccionamos la opción "Riesgo". La sintaxis de todo lo anterior y el resultado es el siguiente:

```
AUTORECODE VARIABLES=EstadAcido Osteop
/INTO Acido Osteoporo
/DESCENDING
/PRINT.

CROSSTABS
/TABLES=Acido BY Osteoporo
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ RISK
/CELLS=COUNT COLUMN
/COUNT ROUND CELL.
```

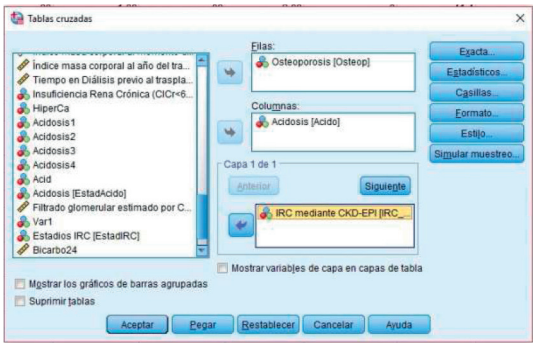
Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Razón de ventajas para Acido (Sí / No)	4,400	1,145	16,913
Para cohorte Osteoporo = Sí	3,217	1,091	9,489
Para cohorte Osteoporo = No	,731	,532	1,006
N de casos válidos	60		

La significación estadística ya la habíamos visto con una p=0.044 según la Prueba exacta de Fisher. Vemos en la tabla anterior los mismos resultados que con los cálculos realizados a

mano: OR=4.4 "Razón de ventajas para Acido (Sí/No)". Y un Riesgo "Para cohorte Osteoporo = Sí" de 3.22. Además mediante este procedimiento podemos obtener los intervalos de confianza del 95%, que vemos que en ninguno de los dos casos engloban el valor 1 y por tanto son significativos. El Riesgo de Osteoporosis entre los pacientes con Acidosis es 3.22 veces superior que entre los pacientes sin Acidosis, oscilando este Riesgo entre 1.1 y 9.5 veces. Se escribiría así: RR 3.22 (IC95%: 1.09 a 9.49).

Es conveniente saber realizar los cálculos del RR "a mano". En el ejemplo anterior hemos utilizado sólo 2 variables (exposición y respuesta), pero en un estudio de seguimiento (prospectivo, cohortes), generalmente vamos a tener en cuenta la presencia de variables de confusión y habrá que "ajustar" por ellas. Ya hemos comentado que esto último se puede hacer mediante Regresión, pero los resultados no son riesgos sino Odds y razones de Odds. Mediante tablas de contingencias el problema crece cuando tenemos que introducir variables de ajuste. Cuantas más variables, más laborioso. Pero es posible hacerlo. Como ejemplo, vamos a introducir la presencia de IRC y calcular el RR de osteoporosis en los pacientes con acidosis; es decir, vamos a ver el riesgo de osteoporosis en los pacientes con acidosis ajustado por la presencia de IRC. El cuadro de diálogo de la tabla de contingencia sería el siguiente:



En el cuadro de la opción "Capa 1 de 1" seleccionamos la variable de ajuste. En este caso "IRC mediante CKD-EPI". Si tenemos más variables de ajuste debemos introducirlas haciendo Clic sobre el botón "Siguiente".

La tabla de contingencia y la sintaxis sería la siguiente:

```
CROSSTABS /TABLES=Osteop BY Acido BY IRC_CKD
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL.
```

Tabla cruzada Osteop*Acido*IRC_CKD

IRC_CKD				Acido		
				Sí	No	Total
No	Osteop	No	Recuento	2	21	23
			% dentro de Acido	50,0%	95,5%	88,5%
		Sí	Recuento	2	1	3
			% dentro de Acido	50,0%	4,5%	11,5%
	Total		Recuento	4	22	26
			% dentro de Acido	100,0%	100,0%	100,0%
Sí	Osteop	No	Recuento	13	12	25
			% dentro de Acido	68,4%	80,0%	73,5%
		Sí	Recuento	6	3	9
			% dentro de Acido	31,6%	20,0%	26,5%
	Total		Recuento	19	15	34
			% dentro de Acido	100,0%	100,0%	100,0%
Total	Osteop	No	Recuento	15	33	48
			% dentro de Acido	65,2%	89,2%	80,0%
		Sí	Recuento	8	4	12
			% dentro de Acido	34,8%	10,8%	20,0%
	Total		Recuento	23	37	60
			% dentro de Acido	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

IRC_CKD		Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
No	Chi-cuadrado de Pearson	6,851[c]	1	,009		
	Corrección de continuidad[b]	3,122	1	,077		
	Razón de verosimilitud	4,916	1	,027		
	Prueba exacta de Fisher				,052	,052
	Asociación lineal por lineal	6,586	1	,010		
	N de casos válidos	26				
Sí	Chi-cuadrado de Pearson	,577[d]	1	,447		
	Corrección de continuidad[b]	,136	1	,713		
	Razón de verosimilitud	,588	1	,443		
	Prueba exacta de Fisher				,697	,360
	Asociación lineal por lineal	,560	1	,454		
	N de casos válidos	34				
Total	Chi-cuadrado de Pearson	5,094[a]	1	,024		
	Corrección de continuidad[b]	3,706	1	,054		
	Razón de verosimilitud	4,980	1	,026		
	Prueba exacta de Fisher				,044	,028
	Asociación lineal por lineal	5,009	1	,025		
	N de casos válidos	60				

a. 1 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 4,60.

b. Sólo se ha calculado para una tabla 2x2

c. 3 casillas (75,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es ,46.

d. 1 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,97.

Vemos que la tabla aparece dividida en filas con los pacientes con y sin IRC. En el cuadro de la derecha vemos que la significación global sigue siendo estadísticamente significativa con una p=0.044 según la Prueba exacta de Fisher de la fila Total; es decir, a pesar de ajusta por la presencia de IRC, la acidosis metabólica aumenta el riesgo de osteoporosis. Pero vemos que la significación es distinta según los pacientes tengan o no IRC, de tal manera que en los pacientes sin IRC no es estadísticamente significativa (aunque casi) con una p= 0.052, mientras que en los pacientes con IRC la significación es claramente no significativa con una p=0.697. Es decir, en los pacientes sin IRC el efecto es más importante que en los pacientes con IRC. De hecho vemos en el cuadro de la izquierda, que en los pacientes sin IRC hay un 50.0% de osteoporosis en los pacientes con acidosis, frente a un 4.5% de osteoporosis en los pacientes sin osteoporosis. En cambio, en los pacientes con IRC, hay un 31.6% de osteoporosis en los pacientes con acidosis frente a un 20.0% de osteoporosis en los pacientes sin acidosis. Vamos a calcular "a mano" los riesgos. Tendremos que calcular dos RR, uno para pacientes con IRC y otro para pacientes sin IRC:

Sin IRC: RR= (2/4)/(1/22) = 11

Con IRC: RR= (6/19)/(3/15) = 1.58

En pacientes sin IRC el riesgo de osteoporosis en los pacientes con acidosis es 11 veces superior frente a los pacientes sin acidosis, mientras que en los pacientes con IRC el riesgo es inferior. No tenemos los intervalos de confianza y por tanto así no nos da toda la información.

En estos cálculos nos faltaría obtener los intervalos de confianza al 95%. Para ello vamos a necesitar una calculadora científica, pero no son difíciles de realizar.

El RR con su intervalo de confianza al 95% se obtiene así:

$$RR \times e^{\pm 1.96 \times SE(lnRR)}$$

Donde SE(lnRR)= $= \sqrt{\frac{1}{a0} - \frac{1}{n0} + \frac{1}{a1} - \frac{1}{n1}}$

a_0= Número de casos en los no expuestos.

n_0= Total de no expuestos.

a_1= Número de casos en los expuestos.

n_1= Total de expuestos.

En nuestro ejemplo SE(lnRR) para los pacientes sin IRC= $= \sqrt{\frac{1}{1} - \frac{1}{22} + \frac{1}{2} - \frac{1}{4}} =$= 1.0975

El IC95% sería: $e^{1.96\times1.0975}$ y $e^{-1.96\times1.0975}$; 8.59 y 0.116. Multiplicado cada uno por el RR sería: 1.28 a 94.54.

Por tanto el RR para pacientes sin IRC sería: 11.00 (IC95%: 1.28 a 94.54).

Para los pacientes con IRC, SE (lnRR) = : $\sqrt{\frac{1}{3} - \frac{1}{15} + \frac{1}{6} - \frac{1}{19}}$ = = 0.6168

El IC95% sería: $e^{1.96\times0.6168}$ y $e^{-1.96\times0.6168}$; 3.347 y 0.299. Multiplicado cada uno por el RR sería: 0.47 a 5.29.

Por tanto el RR para pacientes con IRC sería: 1.58 (IC95%: 0.47 a 5.29). Este intervalo incluye el valor 1 y por tanto ya podemos decir que en los pacientes con IRC el efecto no es significativo.

Vamos a calcularlos con las tablas de contingencia. Debemos proceder de la misma manera que con las tablas de 2 x 2, es decir, recodificar las variables y poner en filas la exposición y en columnas la respuesta. La variable acidosis y osteoporosis ya las teníamos recodificada. Vamos a recodificar la variable IRC con el AUTORECODE; vamos a denominar la nueva variable como "IRCrecod". Tras realizar la tabla de contingencia con las nuevas variables, la sintaxis y el resultado es el siguiente:

```
AUTORECODE  VARIABLES=IRC_CKD
/INTO IRCrecod
/DESCENDING
/PRINT.
CROSSTABS

/TABLES=Acido BY Osteoporo BY IRCrecod
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ RISK
/CELLS=COUNT
/COUNT ROUND CELL.
```

Estimación de riesgo

IRCrecod		Valor	Intervalo de confianza de 95 % Inferior	Superior
Sí	Razón de ventajas para Acido (Sí / No)	1,846	,376	9,077
	Para cohorte Osteoporo = Sí	1,579	,471	5,291
	Para cohorte Osteoporo = No	,855	,575	1,272
	N de casos válidos	34		
No	Razón de ventajas para Acido (Sí / No)	21,000	1,271	346,934
	Para cohorte Osteoporo = Sí	11,000	1,280	94,537
	Para cohorte Osteoporo = No	,524	,196	1,402
	N de casos válidos	26		
Total	Razón de ventajas para Acido (Sí / No)	4,400	1,145	16,913
	Para cohorte Osteoporo = Sí	3,217	1,091	9,489
	Para cohorte Osteoporo = No	,731	,532	1,006
	N de casos válidos	60		

Vemos que los resultados son similares a los realizados "a mano", con un RR de 1.58 para los pacientes con IRC (RR= 1.58, IC95%: 0.47 a 5.29), y de 11.00 para los pacientes sin IRC (RR=11.00, IC95%: 1.28 a 94.54). En la última fila vemos el efecto global de la prueba sin ajustar.

El IC 95% de los pacientes con IRC incluye el valor 1, y por tanto no es significativo como indicaban las pruebas de significación; sin embargo, el IC95% para los pacientes sin IRC no incluye el valor 1, y por tanto sí sería estadísticamente significativo, pero cuando hemos visto la significación estadística el valor que obteníamos mediante la Prueba exacta de Fisher era de p=0.052 (no estadísticamente significativo), aunque el valor de Chi-cuadrado sí lo era con una p=0.009.

Se nos plantea un verdadero problema de decisión. ¿Con cuál nos quedamos? Los intervalos de confianza no están calculados teniendo en cuenta la frecuencia mínima esperada en cada casilla y por tanto no sería aplicable en este caso. El problema reside en que si nos fijamos en la tabla de contingencia, sólo hay 4 pacientes con acidosis en el grupo sin IRC. Es un problema del tamaño de la muestra.

Con más variables de ajuste, la situación se complica aún más, pues deberíamos dar diferentes RR según la categoría de cada variable. Elaborar tablas de contingencia con SPSS con muchas variables puede ser peligroso pues debemos tener la precaución de tener correctamente codificadas las variables. De ahí el saber hacer los cálculos "a mano".

La principal limitación de las tablas de contingencia es que no se pueden introducir variables cuantitativas. Pero una variable cuantitativa siempre se puede categorizar. Por ejemplo si debemos ajustar por la variable "Edad", ésta podemos categorizarla por ejemplo en pacientes "Mayores 65 años" y "Menores 65 años" mediante el procedimiento COMPUTE y así tener una variable binaria. Aunque por otra parte es cierto que categorizar una variable cuantitativa es perder información. Por ello generalmente los ajustes por variables se suelen hacer mediante Regresión pues no precisa de un proceso tan laborioso como el de las tablas de contingencia y nos permite introducir variables cuantitativas tal cual, y aunque los resultados nos los expresa como Odds y razones de Odds (OR), son conceptos igual de interpretables que el RR. Por ello vamos a dejar el cálculo del RR para

los estudios donde los grupos de exposición estén perfectamente balanceados y por tanto no se precisaría (teóricamente) de ajuste por otras variables como por ejemplo en los Ensayos clínicos randomizados. Veremos estos conceptos con más detalle cuando hablemos de regresión más adelante.

CÁLCULO DE NÚMERO NECESARIO A TRATAR (NNT)

El concepto de NNT va ligado a cualquier estudio de intervención que se realice, como los Ensayos Clínicos, y últimamente es un dato que debe aparecer siempre en este tipo de estudios porque expresa de manera bastante clara la magnitud de la intervención. Hoy en día estamos habituados a ver estudios que nos presentan los distintos laboratorios farmacéuticos sobre las bondades de sus productos donde los resultados nos los expresan a través de magníficos gráficos (muchos de ellos editados para engañar) como curvas de supervivencia, y otros datos como los riesgos relativos, etc., inclinándonos a utilizar dichos productos porque por ejemplo nos digan que disminuye el riesgo relativo de muerte un 26%. La mayoría de estos estudios están realizados con un tamaño de muestra muy grande donde cualquier diferencia por muy pequeña que sea, va a ser fácilmente demostrada estadísticamente, y además con características de los pacientes que probablemente no son los que vemos habitualmente en la práctica clínica diaria. Por tanto, la pregunta que nos tenemos que hacer no es ¿cuánto disminuye el riesgo de muerte el producto? sino ¿cuántos pacientes tengo que tratar con dicho producto para que uno de ellos no fallezca? De nada sirve que un fármaco reduce el riesgo de muerte un 26% si para conseguir que uno no fallezca debo tratar a 900, con sus consiguientes efectos secundarios.

Por estos motivos surgió el concepto de NNT. Se interpreta como el número de pacientes que debo tratar para que uno de ellos no desarrolle el evento en cuestión. Se calcula como la inversa de la diferencia de proporciones o riesgos:

$$NNT = \frac{1}{RD}$$

Para el ejemplo de acidosis y osteoporosis donde vimos que en el grupo con acidosis había un 24% más de pacientes con osteoporosis que en el grupo sin acidosis, con una diferencia de riesgo de 0.24, el valor del NNT será: NNT= 1/0.24 = 4.17 (se redondea siempre al entero superior). En este caso representa que por cada 5 pacientes con acidosis, 1 va a desarrollar osteoporosis. Este valor nos puede orientar sobre la importancia de la exposición y la utilidad de un tratamiento.

CÁLCULO DE LA TASA DE INCIDENCIA

Como se ha comentado anteriormente, los pacientes al comienzo de un estudio de seguimiento, van a estar en "riesgo" de desarrollar la enfermedad en cuestión (o el proceso que se esté estudiando). La razón entre el riesgo de los pacientes expuestos frente al de los pacientes no expuestos es lo que hemos denominado Razón de Riesgo, Riesgo Relativo o Razón de Incidencias (RR o RI). Son el mismo concepto. Sin embargo, en los

estudios que conllevan seguimiento, se suele tener en cuenta el factor tiempo porque nos va a dar información sobre la "velocidad" a la que se "producen los enfermos". A esto es a lo que denominamos "Tasa de Incidencia". Veamos el siguiente ejemplo: imaginemos que durante 1 mes damos dos fármacos antihipertensivo a 2 grupos de 20 pacientes cada grupo de hipertensos, y que al final del mes en ambos grupos hay 10 pacientes a los que se les controla la tensión arterial. En ambos grupos la incidencia (o riesgo) de control de la tensión arterial es del 50% y por tanto el RR sería de 1; podríamos pensar que ambos fármacos son iguales. Pero imaginemos ahora que en el primer grupo el control de la tensión arterial se consiguió ya durante la primera semana y que en el segundo grupo se consiguió al final de la última semana del mes de seguimiento. El riesgo sigue siendo del 50% en ambos grupos al finalizar el mes, sin embargo podemos intuir que el fármaco del primer grupo tiene una eficacia más rápida que la del segundo grupo. Esta información la aporta la tasa de incidencia.

Vamos a calcular la tasa en ambos grupos. Para ello debemos tener en cuenta el tiempo que ha estado en "riesgo de controlar la tensión arterial" cada paciente de grupo. En el primer grupo 10 pacientes habrán estado en riesgo todo el mes pues no se ha controlado la tensión arterial (por tanto 10 x 30 días) y otros 10 habrán estado en riesgo sólo la primera semana, pues después se controló la tensión (por tanto 10 x 7 días). En el segundo grupo, 10 pacientes han estado en riesgo todo el mes al no controlarse la tensión al finalizar el seguimiento (por tanto 10 x 30 días), mientras que los otros 10 habrán estado en riesgo sólo las 3 primeras semanas pues en la última se controló (por tanto 10 x 21 días). Tenemos:

Grupo 1: $10/(10 \times 30) + (10 \times 7) = 0.0270$

Grupo 2: $10/(10 \times 30) + (10 \times 21) = 0.0196$

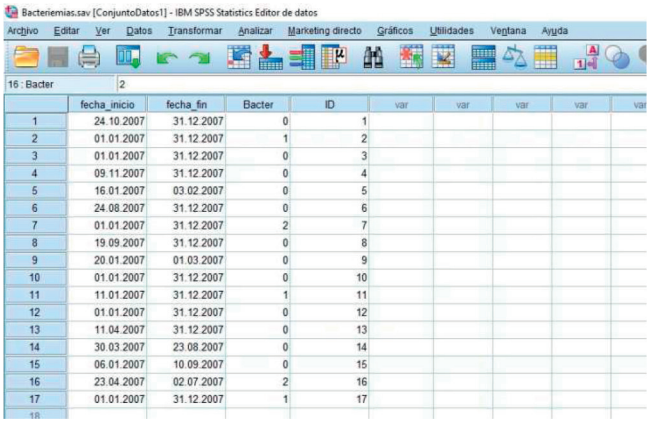
En el grupo 1 se producen 0.0270 controles de la tensión arterial por cada paciente y día, mientras que en el grupo 2 se producen 0.0196 controles de tensión por cada paciente y día. Si multiplicamos ambos valores por ejemplo por 100, en el grupo 1 tenemos que se producen 2.7 controles de tensión arterial por cada 100 pacientes y día (2.7 controles/100 pacientes-día) y en el grupo 2 redondeando se producen 2.0 controles por cada 100 pacientes y día (2.0 controles/100 pacientes-día).

La razón entre ambas tasas de incidencias es lo que se conoce como Hazard Ratio que veremos en los estudios de supervivencia. En este ejemplo el valor del HR es 1.38, indicando superioridad del primer fármaco frente al segundo.

Para hacer los cálculos anteriores con SPSS necesitamos en nuestra base de datos 3 variables: una con la fecha de inicio del estudio (la fecha en la que el paciente entre en el estudio; hay pocos estudios en los que los pacientes comienzan todos desde el principio, lo normal es que vayan entrando progresivamente a medida que van apareciendo); una variable con la fecha fin del estudio (es la fecha en la que se acaba el estudio, o la fecha en la que el paciente sale del estudio porque haya presentado el evento); y una variable que recoja el estado en el que se encuentra el paciente en la fecha que sale del estudio (esta variable será 0 si no ha presentado el evento, 1 si lo ha presentado, 2

si se ha perdido el seguimiento sin haber presentado el evento, 3 otro motivo…). Los pacientes que se pierden en el seguimiento sin haber presentado el evento también entran a formar parte del tiempo de seguimiento porque son tiempos en los que ha estado en riesgo y no han presentado el evento. Si lo que analizamos es el evento "muerte por una causa", al final del seguimiento el 0 serán los vivos, 1 los muertos, 2 las pérdidas de seguimiento pero vivos y 3 muertes por otras causas (lo que denominamos muertes censuradas). Podemos poner todas las categorías que estimemos oportunas siempre que quede bien definida cual es la categoría del evento (1= muertos).

Para poner un ejemplo de lo anterior y ver como se realiza con SPSS vamos a utilizar la base de datos "Bacteriemias.sav", que vemos a continuación:


Bacteriemias.sav [ConjuntoDatos1] - IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

16 : Bacter 2

	fecha_inicio	fecha_fin	Bacter	ID	var	var	var	var	var
1	24.10.2007	31.12.2007	0	1					
2	01.01.2007	31.12.2007	1	2					
3	01.01.2007	31.12.2007	0	3					
4	09.11.2007	31.12.2007	0	4					
5	16.01.2007	03.02.2007	0	5					
6	24.08.2007	31.12.2007	0	6					
7	01.01.2007	31.12.2007	2	7					
8	19.09.2007	31.12.2007	0	8					
9	20.01.2007	01.03.2007	0	9					
10	01.01.2007	31.12.2007	0	10					
11	11.01.2007	31.12.2007	1	11					
12	01.01.2007	31.12.2007	0	12					
13	11.04.2007	31.12.2007	0	13					
14	30.03.2007	23.08.2007	0	14					
15	06.01.2007	10.09.2007	0	15					
16	23.04.2007	02.07.2007	2	16					
17	01.01.2007	31.12.2007	1	17					

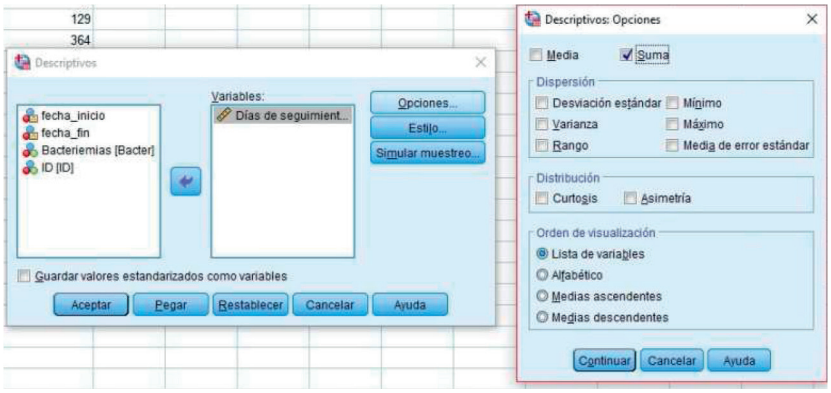
Se trata de un estudio en el que queremos valorar la tasa de incidencias de bacteriemias por catéter venoso central en una población de 17 pacientes en diálisis durante 1 año con fecha del final del mismo el 31/12/2007. Vemos que tenemos las variables "fecha de inicio" que es el momento en el que el paciente comienza diálisis a través de catéter; "fecha fin" como la fecha fin del estudio, la fecha en la que el paciente tiene su primera bacteriemia o la fecha en la que se pierde el seguimiento. Y la variable "Bacter" que indica la situación de cada paciente al finalizar el estudio: 0 sin bacteriemia, 1 con bacteriemia y 2 pérdida de seguimiento. En primer lugar debemos calcular el número de días que se ha seguido a cada paciente a través de las fechas fin e inicio. La sintaxis sería:

COMPUTE dias=CTIME.DAYS(fecha_fin - fecha_inicio).

EXECUTE.

Se nos creará la variable "días" que debemos definir. A continuación calcularemos el tiempo total de seguimiento de la muestra con la suma de los días que cada paciente ha estado en riesgo de padecer la bacteriemia. El procedimiento con SPSS es el siguiente:

Analizar → Estadísticos descriptivos → Descriptivo. Se nos abrirá el siguiente cuadro de diálogo:



Seleccionamos la variable "Días de seguimiento" y la desplazamos hacia el cuadro de la derecha "Variables". Al hacer Clic sobre el botón "Opciones" se nos abre el cuadro de diálogo de la derecha y seleccionamos la opción "Suma"; de esta manera nos sumará los días en riesgo de todos los pacientes. Tras ejecutar la sintaxis vemos que la suma total es de 3675 días. Estos son los días de riesgo totales. Por otra parte debemos ver cuantos pacientes han presentado bacteriemias. Para ello lo realizamos a través de la opción de FREQUENCIES. El procedimiento es el siguiente:

Analizar → Estadísticos descriptivos → Frecuencias. Seleccionamos la variable "Bacter". Vemos que se han producido 3 bacteriemias.

Por tanto la tasa de incidencia sería: 3/3675 = 0.008. Es decir, se producen 0.008 bacteriemias por cada paciente y día. Si multiplicamos por 1000 diremos que se producen 8 bacteriemias por cada 1000 pacientes y día (8 bacteriemias/1000 pacientes - día).

ANÁLISIS ESTADÍSTICO BÁSICO PARA ESTUDIOS CON MEDIDAS EN UN MISMO SUJETO

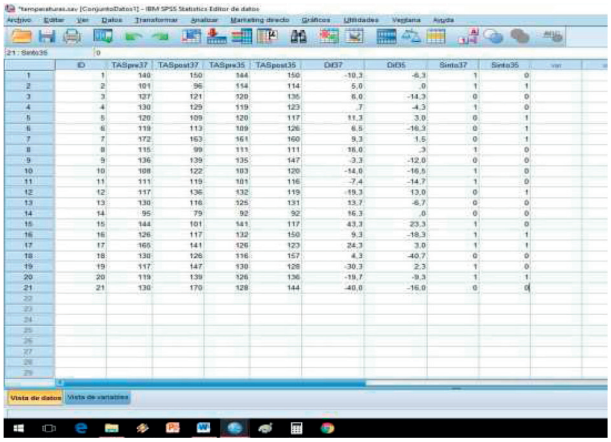
Hasta ahora el análisis estadístico se ha realizado comparando variables categóricas o cuantitativas en 2 o más grupos, pero en determinados estudios no vamos a tener un grupo control con el que comparar sino que sólo vamos a tener un grupo, donde un mismo paciente va a ser el control de sí mismo. Este tipo de análisis es frecuente en los ensayos clínicos. Por ejemplo, queremos comparar el efecto de un fármaco A para bajar la tensión arterial frente a placebo; en un primero momento, de manera aleatoria, a un grupo se le administra el fármaco A y a otro el placebo, y al cabo de las semanas al primer grupo se le cambia a placebo y al segundo se le administra el fármaco A. Son los denominados estudios cruzados (Cross-over study).

La principal limitación de este tipo de estudios es la influencia del "efecto periodo". Durante el periodo del estudio pueden suceder diversas circunstancias que hagan que los periodos en los que se realiza la intervención no sean comparables. Por ejemplo imaginemos que comenzamos dando el fármaco A a todos los pacientes y se consigue una reducción de la tensión sistólica de 20 mmHg, y posteriormente damos el placebo a todos los pacientes y se consigue una reducción de la tensión de tan sólo 5 mmHg.

¿Podemos concluir que el fármaco A es mejor que el placebo? Imaginemos que el periodo en el que se administra el fármaco A es verano, y que es invierno en el periodo en el que damos el placebo. ¿Qué parte del descenso de la tensión arterial con el fármaco A no es debida al efecto hipotensor del calor del verano?

Este efecto periodo se consigue neutralizar generalmente con los estudios cruzados, donde en un primer periodo damos el fármaco A y el placebo al azar entre los pacientes, y posteriormente se cruzan las actuaciones en el segundo periodo.

Para el ejemplo de este tipo de análisis con SPSS vamos a utilizar la base de datos "Temperaturas.sav". Queremos ver si el descenso de la tensión arterial en los pacientes en diálisis al comienzo y al final de la sesión es diferente según la temperatura con la que se programe la máquina de diálisis. Los pacientes se han asignado de manera aleatoria a recibir una sesión de diálisis con 37ºC o con 35ºC en el primer periodo y posteriormente se han cruzado cambiando las temperaturas. La base de datos es la siguiente:



	ID	TASpre37	TASpost37	TASpre35	TASpost35	Dif37	Dif35	Sinto37	Sinto35	var
1	1	140	150	144	150	-10,3	-6,3	1	0	
2	2	101	96	114	114	5,0	,0	1	1	
3	3	127	121	120	135	6,0	-14,3	0	0	
4	4	130	129	119	123	,7	-4,3	1	0	
5	5	120	109	120	117	11,3	3,0	0	1	
6	6	119	113	109	126	6,5	-16,3	0	1	
7	7	172	163	161	160	9,3	1,5	0	1	
8	8	115	99	111	111	16,0	,3	1	0	
9	9	136	139	135	147	-3,3	-12,0	0	0	
10	10	108	122	103	120	-14,0	-16,5	1	0	
11	11	111	119	101	116	-7,4	-14,7	1	0	
12	12	117	136	132	119	-19,3	13,0	0	1	
13	13	130	116	125	131	13,7	-6,7	0	0	
14	14	95	79	92	92	16,3	,0	0	0	
15	15	144	101	141	117	43,3	23,3	1	0	
16	16	126	117	132	150	9,3	-18,3	1	1	
17	17	165	141	126	123	24,3	3,0	1	1	
18	18	130	126	116	157	4,3	-40,7	0	0	
19	19	117	147	130	128	-30,3	2,3	1	0	
20	20	119	139	126	136	-19,7	-9,3	1	1	
21	21	130	170	128	144	-40,0	-16,0	0	0	

Tenemos las tensiones registradas al comenzar y al finalizar la sesión de diálisis con ambas temperaturas, así como la presencia de síntomas con una u otra temperatura.

Para el análisis, en primer lugar vamos a ver la diferencia de caída de la tensión arterial sistólica con ambas temperaturas, es decir, vamos a calcular las diferencias. La sintaxis sería:

```
COMPUTE Dif37=TASpre37 - TASpost37.
COMPUTE Dif35=TASpre35 - TASpost35.
EXECUTE.

* Definir propiedades de variables.
*Dif37.
VARIABLE LABELS Dif37 'Variación de TAS con 37ºC'. FORMATS Dif37(F5.1).
*Dif35.
VARIABLE LABELS Dif35 'Variación de TAS con 35ºC'. FORMATS Dif35(F5.1).
EXECUTE.
```

En segundo lugar debemos comprobar si las variables creadas siguen o no una distribución normal.

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Dif37	,140	21	,200*	,974	21	,827
Dif35	,149	21	,200*	,946	21	,284

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

a. Corrección de significación de Lilliefors

Comprobamos que las pruebas de normalidad no son significativas y por tanto podemos asumir que siguen una distribución Normal.

A continuación comparamos ambos cambios de tensión arterial con el siguiente procedimiento:

Analizar → Comparar medias → Pruebas T para muestras relacionadas. Se nos abrirá el siguiente cuadro de diálogo:

En las casillas de "Variables emparejadas" seleccionamos los pares de variables a comparar.

En este ejemplo sólo tenemos un par de variables a comparar. La sintaxis y la tabla resumen es la siguiente:

T-TEST PAIRS=Dif37 WITH Dif35 (PAIRED)
/CRITERIA=CI(.9500)
/MISSING=ANALYSIS.

Estadísticas de muestras emparejadas

		Media	N	Desviación estándar	Media de error estándar
Par 1	Dif37	1,036	21	19,0052	4,1473
	Dif35	-6,143	21	13,2052	2,8816

Correlaciones de muestras emparejadas

		N	Correlación	Sig.
Par 1	Dif37 & Dif35	21	,316	,163

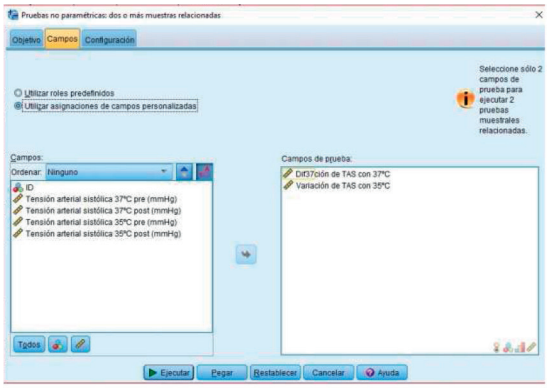
Prueba de muestras emparejadas

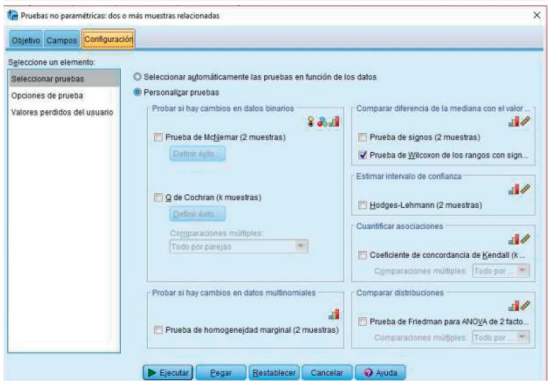
		Diferencias emparejadas					t	gl	Sig. (bilateral)
		Media	Desviación estándar	Media de error estándar	95% de intervalo de confianza de la diferencia				
					Inferior	Superior			
Par 1	Dif37 - Dif35	7,1795	19,4162	4,2370	-1,6586	16,0177	1,694	20	,106

En la primera tabla aparecen las medias y las desviaciones estándar de cada variable. En la segunda tabla aparece la correlación entre ambas. Vemos que tienen una correlación con R=0.316, pero no es significativa (p=0.163). En la última tabla vemos la significación de la comparación. La media de las diferencias es de 7.18 mmHg con un IC95% entre -1.66 y 16.02. Como incluye el valor 0 implica que ambas variables son iguales y por eso la diferencia es 0. Por ello vemos que no es significativa la diferencia con una p=0.106.

Si las variables no siguen una distribución Normal, el procedimiento es el siguiente:

Analizar → Pruebas no paramétricas → Muestras relacionadas. Se nos abrirán los siguientes cuadros de diálogo:





En la pestaña "Campos" seleccionamos las 2 variables a comparar y las desplazamos hacia la derecha. En la pestaña "Configuración" seleccionamos la opción "Personalizar pruebas" y posteriormente marcamos la opción "Prueba de Wilcoxon de los rangos con sign…". La sintaxis y el resultado sería la siguiente:

```
NPTESTS
/RELATED TEST(Dif37 Dif35) WILCOXON
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.
```

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La mediana de las diferencias entre Variación de TAS con 37ºC y Variación de TAS con 35ºC es igual a 0.	Prueba de rangos con signo de Wilcoxon para muestras relacionadas	,092	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

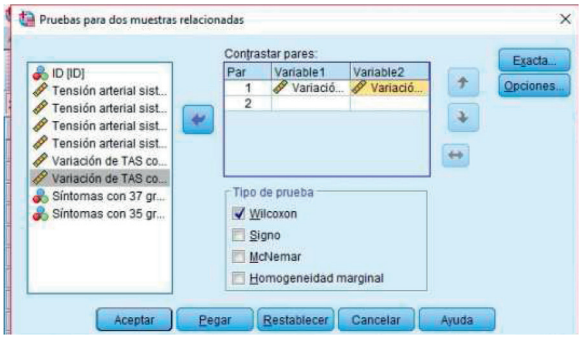
Vemos que la prueba es no significativa con p=0.092, y la tabla nos indica conservar la hipótesis nula, es decir, la variación de tensión arterial con ambas temperaturas es la misma.

Si hacemos doble Clic sobre la tabla anterior obtendremos una nueva ventana de resultados con el resumen de la prueba.

En caso que no funcione el procedimiento anterior, debemos hacer lo siguiente:

Analizar → Pruebas no paramétricas → Cuadros de diálogos antiguos → 2 muestras relacionadas.

Se nos abrirá el siguiente cuadro de diálogo:



Seleccionamos las 2 variables a estudio y en "Tipo de prueba" seleccionamos la opción Wilcoxon.

El resultado y la sintaxis son las siguientes:

```
NPAR TESTS
/WILCOXON=Dif37 WITH Dif35 (PAIRED)
/MISSING ANALYSIS.
```

Estadísticos de prueba[a]

	Dif35 - Dif37
Z	-1,686[b]
Sig. asintótica (bilateral)	,092

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos positivos.

Vemos que el resultado es igual al anterior.

En el ejemplo anterior se ha realizado con variables cuantitativas. Para el ejemplo con variables categóricas vamos a utilizar la presencia de síntomas con una u otra temperatura.

La tabla de contingencia con ambas variables sería la siguiente:

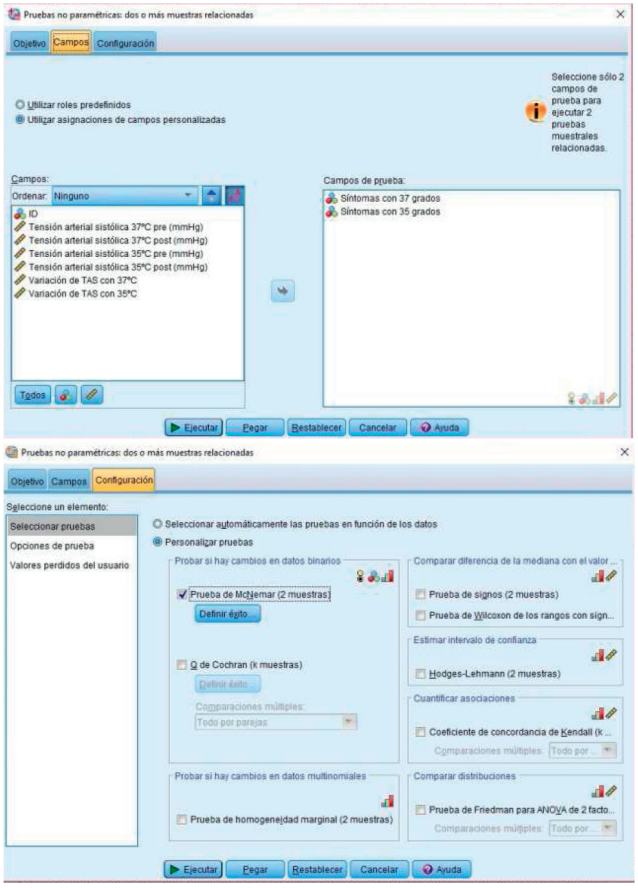
Tabla cruzada Sinto37*Sinto35

Recuento

		Sinto35		
		0	1	Total
Sinto37	0	6	4	10
	1	7	4	11
Total		13	8	21

En esta tabla vemos que hay pacientes que no presentan síntomas con ninguna de las 2 temperaturas y pacientes que desarrollan síntomas con las 2 temperaturas. Por otra parte hay pacientes que desarrollan síntomas con una temperatura y con otra no, que son los pacientes de la diagonal. Estos son los pacientes a tener en cuenta para valorar las diferencias. El procedimiento es el siguiente:

Analizar → Pruebas no paramétricas → Muestras relacionadas. Se nos abrirán los siguientes cuadros de diálogo:



Seleccionamos las 2 variables a comparar, y en la pestaña "Configuración" en esta ocasión seleccionamos la "Prueba de McNemar" (2 muestras).

La sintaxis y la tabla de resultado es la siguiente:

```
NPTESTS
/RELATED TEST(Sinto37 Sinto35) MCNEMAR(SUCCESS=FIRST)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.
```

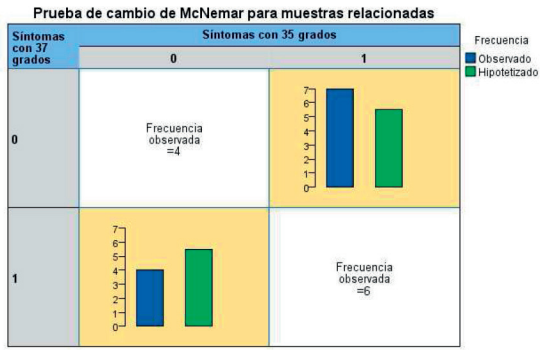
Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	Las distribuciones de valores diferentes entre Síntomas con 37 grados y Síntomas con 35 grados tienen las mismas probabilidades.	Prueba de McNemar para muestras relacionadas	,549[1]	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

[1] Se muestra la significación exacta para esta prueba.

Vemos que no es significativo con una p=0.549, es decir, la proporción de síntomas es igual para ambas temperaturas. Si hacemos doble Clic sobre la tabla anterior se nos abrirá una nueva ventana con el resumen del procedimiento:



Vemos en las casillas sombreadas de amarillo que sólo se tiene en cuenta las casillas de la diagonal como habías indicado anteriormente.

En caso que no funcione el procedimiento anterior, debemos hacer lo siguiente:

Analizar → Pruebas no paramétricas → Cuadros de diálogos antiguos → 2 muestras relacionadas.

Se nos abrirá un cuadro de diálogo similar al anterior:

Seleccionamos las 2 variables categóricas a comparar y seleccionamos en esta ocasión en "Tipo de prueba" la opción McNemar. La sintaxis y resultados son los siguientes:

NPAR TESTS
/MCNEMAR=Sinto37 WITH Sinto35 (PAIRED)
/MISSING ANALYSIS.

Estadísticos de prueba[a]

	Sinto37 & Sinto35
N	21
Significación exacta (bilateral)	,549[b]

a. Prueba de McNemar

b. Distribución binomial utilizada.

Vemos que el resultado es igual que como se ha realizado anteriormente.

ANÁLISIS ESTADÍSTICO AVANZADO. MODELOS DE REGRESIÓN

En los temas anteriores hemos visto los procedimientos estadísticos necesarios para comparar 2 variables: una variable dependiente (o respuesta), la cual podía ser cuantitativa o categórica, y una variable independiente (variable exposición o predictora) que podía tener 2 o más categorías. Según la combinación entre ellas teníamos los procedimientos estadísticos resumidos en la siguiente tabla:

	Variable exposición o predictora		
Variable respuesta	**2 categorías**	**> 2 categorías**	**Cuantitativa**
Cuantitativa	• T-Student • Pruebas no paramétricas • **Regresión Lineal**	• ANOVA • Pruebas no paramétricas • **Regresión Lineal**	• Correlación • **Regresión Lineal**
Categórica	• Chi-cuadrado • Estadístico Exacto de Fisher • **Regresión Logística** • **Regresión Cox**	• Chi-Cuadrado • Prueba Tendencia Lineal • **Regresión Logística** • **Regresión Cox**	• **Regresión Logística** • **Regresión Cox**

Como vemos en la tabla anterior, se han introducido Modelos de Regresión que nos permiten realizar cualquiera de los análisis que hemos visto hasta ahora.

Los métodos estadísticos vistos hasta ahora permiten tan sólo la comparación entre 1 variable exposición y una variable respuesta, pero esta situación va a suceder muy pocas veces.

En la recogida de datos para un estudio, el azar puede hacer que el muestreo provoque desajustes en determinadas variables entre los grupos a comparar que influyan en los resultados observados. Por ejemplo, en los resultados anteriores hemos visto que la presencia de acidosis aumenta el riesgo de desarrollar osteoporosis. Pero es conocido por otros estudios que los corticoides también aumentan el riesgo de osteoporosis. Imaginemos que la dosis de corticoides entre los pacientes con acidosis es el doble que en los pacientes sin acidosis. ¿Qué parte del riesgo de osteoporosis es debida a la mayor dosis de corticoides y no a la propia

presencia de acidosis? Para ello debemos "extraer" primero el efecto de los corticoides sobre el desarrollo de osteoporosis para quedarnos con el efecto "neto" de la acidosis. Esto es lo que se conoce como "ajuste por otras variables".

El procedimiento estadístico por el que se puede ver el efecto de una variable exposición sobre una variable respuesta ajustado por otras variables se realiza mediante los modelos de regresión.

Vamos a ver 3 tipos de modelos de regresión, los cuales van a depender del tipo de variable respuesta:

- Respuesta cuantitativa: Regresión Lineal.
- Respuesta categórica: Regresión Logística.
- Respuesta categórica con factor tiempo: Regresión de Cox. Análisis de la supervivencia.

Vemos que no hemos hecho ninguna alusión al tipo de variable independiente, la cual podrá ser cuantitativa o categórica con 2 o más variables.

Cuando en el modelo de regresión tan sólo se tiene en cuenta la variable respuesta y una sola variable independiente se denomina "análisis univariante", y sus resultados van a coincidir con los obtenidos mediante los procedimientos de Chi-Cuadrado o T- Student como luego veremos. Cuando en el modelo de regresión se tienen en cuenta más de una variable independiente se denomina "análisis multivariante".

Los análisis mediante regresión van a tener diferentes finalidades dependiendo del tipo de estudio que estemos realizando:

- **En los estudios prospectivos,** (Estudios de Cohortes, Ensayos Clínicos) la finalidad de los mismos va a ser ver la influencia de una determinada exposición sobre el desarrollo de un determinado evento (desarrollo de enfermedad, eficacia de un tratamiento, etc.). Los pacientes se van a elegir según presenten o no la exposición. A grandes rasgos podemos decir que partimos del presente para ver qué sucede en el futuro (aunque también hay estudios de Cohortes retrospectivas). En estos estudios debemos registrar además otras variables que se hayan demostrado que puedan influir en el desarrollo del evento (variables de confusión). El muestreo puede dar lugar a disbalances en estas variables de confusión entre los grupos a comparar y por tanto habrá que ajustar los resultados a los disbalances observados. Nuestro ejemplo de estudio es un estudio de Cohortes (partimos del desarrollo de acidosis para ver la aparición de osteoporosis) y por tanto debemos ajustar los resultados a las variables que presenten disbalances (por ejemplo ajustar por la variable corticoides si las dosis son distintas entre los pacientes con o sin acidosis). En los Ensayos Clínicos la aleatorización de los pacientes (y el tamaño de la muestra) precisamente se realiza para evitar la existencia de disbalances estadísticamente significativos entre los grupos a comparar.

En estos estudios la finalidad de la Regresión es la **medición de un efecto ajustado por otras variables.**

- **En los estudios retrospectivos** (estudios Casos-Controles) el objetivo va a ser diferente. Los pacientes se van a elegir según presenten o no el evento y se revisarán antecedentes previos para ver cuáles de ellos pueden haber influido en el desarrollo del evento. Es decir, partimos del evento y tratamos de ver cuáles pueden haber sido las causas que lo han motivado. Si nuestro ejemplo fuese un estudio de casos-controles, los pacientes se hubiesen elegido según presentasen o no osteoporosis, y revisaríamos sus historias clínicas para ver qué antecedentes en el pasado pueden haber influido en el desarrollo de osteoporosis en el presente. El problema de este tipo de estudio es que debemos confiar en lo que hay recogido en las historias clínicas. Cuando alguien realiza una historia clínica no está pensando en que alguien más adelante la pueda necesitar para realizar un estudio y por tanto "el cuidado" con el que se introducen los datos es menor; además es posible que no estén recogidas todas la variables que se necesitan. El que un paciente no tenga registrado la condición de fumador o no, no significa que el paciente sea "No fumador" (puede que quien introdujese los datos olvidase registrar este hecho).

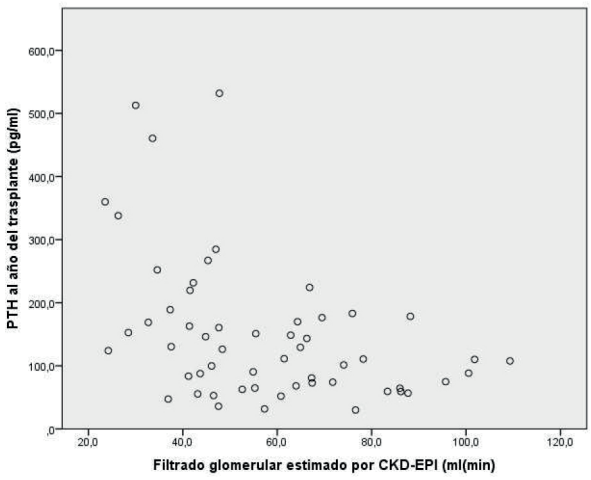
En este tipo de estudio la finalidad de la Regresión va a ser ver cuáles son las variables que determinan la aparición de un efecto: **finalidad descriptiva**; o predecir lo que pueda suceder en el futuro: **finalidad predictiva**.

A pesar de que la finalidad va a ser distinta según el tipo de estudio, el procedimiento del análisis de regresión va a ser el mismo. SPSS no sabe si estamos analizando un estudio de casos-controles o un ensayo clínico y por tanto seremos nosotros los que tenemos que guiar al programa para que realice los análisis que nosotros queremos.

REGRESIÓN LINEAL

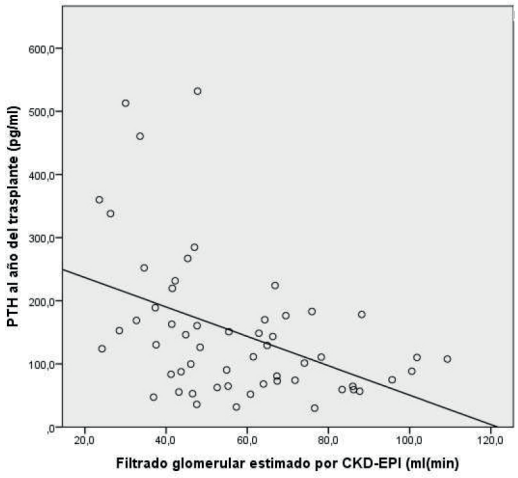
En los modelos de regresión lineal la variable dependiente va a ser cuantitativa.

Para ilustrar este modelo estadístico veamos el siguiente gráfico de dispersión donde se representa los niveles de "PTH al año" en el eje Y, y la variable "Filtrado Glomerular por CKD-EPI" en el eje X:



Aunque el gráfico es similar al que realizamos cuando hablamos de Correlación, aquí sólo tiene sentido ver la relación de PTH en función de CKD-EPI y no a la inversa (no tiene sentido ver la relación de CKD-EPI en función de PTH). Es decir, es una relación asimétrica, mientras que en el estudio de Correlación la relación era simétrica. Se podían extraer dos estudios de correlación, uno de Y en función de X, y otro de X en función de Y. En el estudio de Correlación no nos planteábamos qué variable era la dependiente y cuál era la independiente. En la regresión sin embargo sí tenemos perfectamente definido qué variable es cada cual.

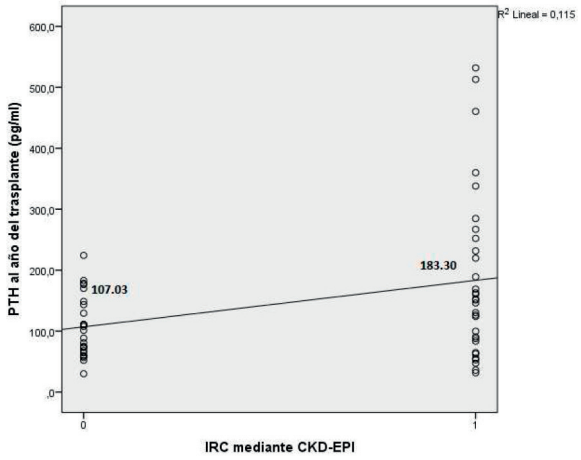
En este gráfico vemos una nube de puntos, pero ¿de qué manera podemos ver la relación entre ambas variables? Se demuestra que la mejor manera de relacionar ambas variables es ajustando una recta a esta nube de puntos. Pero no cualquier recta, sino aquella que mejor "ajuste" a los datos. A esta recta se le denomina "Recta de Regresión", representada en el siguiente gráfico:



Recordando un poco las matemáticas, la ecuación de esta recta es: Y= a + bx, donde el término "a" es el valor de la "constante" donde la recta corta al eje Y (es decir, cuando X=0), y "b" es la pendiente de la recta. Se denomina Regresión Lineal porque la recta de ajuste debe presentar una tendencia lineal, la cual viene determinada por el coeficiente "b". Si no existe tendencia lineal, el coeficiente "b" será igual a cero dando como resultado una recta horizontal situada a la altura del coeficiente "a". Por tanto el valor de "b" será distinto de cero, siendo positivo si la relación es positiva (a medida que aumenta X también aumenta de media Y), o negativo (a medida que aumenta X disminuye de media Y, como en este caso).

La distancia desde cada punto de la nube hasta la recta de regresión es los que se denomina "Residual", que serán importantes posteriormente para ver los diagnósticos de estos modelos. Se demuestra que la recta que ajusta de manera más satisfactoria a los datos es aquella cuya suma al cuadrado de estos residuales es menor. Cuanto más cerca estén los puntos a la recta de regresión menor serán los residuales y por tanto mejor ajustará.

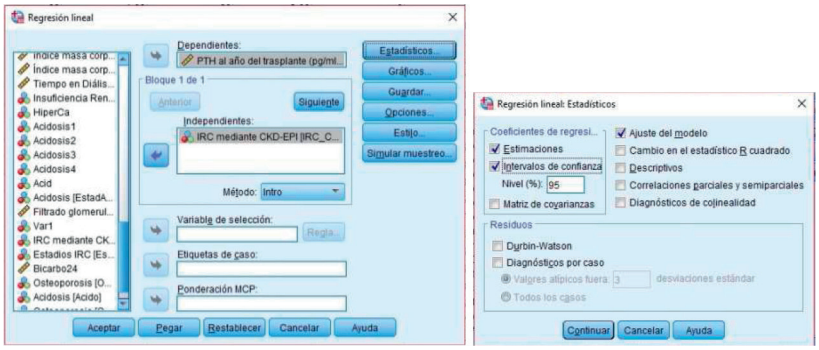
Para ver el significado de cada término de la ecuación de la recta vamos a utilizar como ejemplo el caso más sencillo donde la variable independiente va a ser una variable categórica binaria. Vamos a ver la relación entre la presencia o no de IRC y los valores de PTH al año. El gráfico con su recta de regresión es el siguiente:



En el caso de una variable independiente categórica binaria, se demuestra que la mejor recta es aquella que pasa por la media de cada categoría. Vemos que cuando la categoría de X=0 (Sin IRC), el valor de PTH es 107.03; este valor corresponde al coeficiente "a" de la ecuación. Cuando los pacientes presentan IRC (es decir, X=1) el valor de PTH es 183.30. Existe un aumento de 73.27 en los valores de PTH. Dicho de otro modo, pasar de 0 a 1 en el eje X implica un aumento de 73.27 en los valores de PTH. Este incremento corresponde a la pendiente de la recta de regresión representada por el coeficiente "b". Por tanto, en este caso la ecuación de la recta sería: PTH=107.03 +73.27*X. En este caso el valor del coeficiente "b" es positivo indicando que por cada aumento de X, aumenta de media también Y.

Estos coeficientes son los que se obtienen mediante el análisis de regresión. Veamos cómo se realiza con SPSS. El procedimiento es el siguiente:

Analiza → Regresión → Lineales… Se nos abren los siguientes cuadros de diálogo:



En el primer cuadro seleccionamos la variable dependiente, en este caso "PTH al año" y la arrastramos hasta el cuadro donde pone "Dependientes". Como variable "Independiente" seleccionamos la variable predictora, en este caso "IRC mediante CKD- EPI". Hacemos posteriormente Clic sobre el botón "Estadísticos" y se nos abre el cuadro de diálogo de la derecha. Seleccionamos la opción "Estimaciones" e "Intervalos de confianza". La sintaxis completa sería la siguiente:

REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT PTH4
/METHOD=ENTER IRC_CKD.

Tras ejecutarla tenemos los siguientes resultados:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,338[a]	,115	,098	107,3575

a. Predictores: (Constante), IRC_CKD

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	80508,910	1	80508,910	6,985	,011[b]
	Residuo	622384,330	54	11525,636		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	107,028	21,472		4,985	,000	63,980	150,076
	IRC_CKD	76,272	28,859	,338	2,643	,011	18,414	134,130

a. Variable dependiente: PTH4

Veamos estos resultados. ¿Cómo se valora el grado de ajuste de la recta estimada? Pues viendo la parte explicada por la recta de regresión sobre la variabilidad total de los datos. En la segunda tabla "ANOVA" viene detallada la "Suma de cuadrados" de la variabilidad "Total" de los datos y la de la "Regresión". Si dividimos la suma de cuadrados de la regresión sobre la total obtenemos el "Coeficiente de determinación R2", que viene representado en la primera tabla "Resumen del modelo" como "R cuadrado" cuyo valor es 0.115. Este valor oscila entre 0, cuando la recta no explica nada de la variabilidad total, y 1 cuando explica el 100% de la variabilidad total. Si la recta explicase el 100% de la variabilidad, el coeficiente de determinación valdría 1 y todos los puntos estarían situados sobre la recta. En este caso, la presencia de IRC explica el 11.5% de la variabilidad total de la PTH. ¿Este porcentaje explicado es significativo? Para ello vemos la significación del modelo en la segunda tabla, donde se analiza esta significación mediante una prueba ANOVA; en este

caso la recta ajusta de manera significativa con una p=0.011. En la tercera tabla aparecen los coeficientes de la recta, donde el coeficiente "a" vale 107.03 y el coeficiente "b" vale 76.27. Vemos que obtenemos un valor diferente al que hemos obtenido previamente porque al realizar la regresión no se han tenido en cuenta los valores de PTH perdidos. Por tanto la ecuación de la recta sería: PTH= 107.03 + 76.27*X. Pasar de no tener IRC (X=0) a tener IRC (X=1) supone un aumento de 76.27 los valores de PTH. ¿Este aumento de los valores de PTH es significativo? Para ello vemos el IC 95% de esta pendiente. Si engloba el valor 0 implica ausencia de linealidad. En este caso el IC95% no engloba el valor 0, siendo entre 18.41 y 134.13, y por eso es estadísticamente significativo con una p=0.011.

Por tanto en este ejemplo, pasar de no tener IRC a tenerla implica un aumento de PTH que de media es 76.27, oscilando este aumento entre 18.41 y 134.13. B=76.27, IC95%: 18.41 y 134.13.

La significación del término constante "a" no se tiene en cuenta.

En el caso de una sola variable independiente binaria como en el ejemplo anterior, los resultados coinciden con el de comparación de medias mediante t-Student. En efecto, al realizar este análisis obtenemos:

Estadísticas de grupo

	IRC_CKD	N	Media	Desviación estándar	Media de error estándar
PTH4	No	25	107,028	50,2685	10,0537
	Sí	31	183,300	136,8379	24,5768

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
PTH4	Se asumen varianzas iguales	13,091	,001	-2,643	54	,011	-76,2720	28,8586	-134,1301	-18,4139
	No se asumen varianzas iguales			-2,872	39,498	,007	-76,2720	26,5537	-129,9602	-22,5838

Los valores de PTH para pacientes sin IRC coinciden con el término constante "a" (107.03), mientras que pasar a tener IRC supone un aumento hasta el valor de PTH de 183.30, lo que supone una diferencia de medias de 76.27, que corresponde al término "b". Vemos que la significación de la diferencia y el IC95% son idénticos aunque con signo negativo porque aquí están restando 107.03-183.30.

En el caso de una variable independiente cuantitativa la interpretación es igual que la anterior. Veamos los resultados al comparar los valores de PTH al año en función del Filtrado Glomerular mediante CKD-EPI que fue el ejemplo con el que empezamos este apartado. Los resultados serían:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,436[a]	,190	,175	102,6526

a. Predictores: (Constante), CKD_EPI

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	133865,029	1	133865,029	12,704	,001[b]
	Residuo	569028,211	54	10537,559		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), CKD_EPI

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	283,376	40,054		7,075	,000	203,074	363,679
	CKD_EPI	-2,328	,653	-,436	-3,564	,001	-3,638	-1,019

a. Variable dependiente: PTH4

En este caso la recta explica el 19% de la variabilidad de los valores de PTH, siendo un ajuste estadísticamente significativo, con p=0.001. El valor del coeficiente "b" es -2.33, lo que significa una relación negativa (como vimos en el primer gráfico de este apartado), es decir, por cada unidad de aumento de X implica descenso medio de Y. Aquí, por cada ml/min que aumenta el filtrado glomerular (X) se produce un descenso de PTH que de media es 2.33, oscilando este descenso entre 1.02 y 3.64, siendo estadísticamente significativo, con p=0.001.

Hasta ahora hemos visto los ejemplos con variables independientes cuantitativas y categóricas binarias pero, ¿cómo sería en caso de variables independientes categóricas con más de 2 categorías? Para este tipo de análisis debemos dividir la variable independiente en subcategorías mediante contrastes igual que hicimos con el ANOVA. El problema es que SPSS no sabe cómo hacerlo y hay que hacerlo "a mano". Vamos a explicar el ejemplo utilizando una categoría de referencia, que suele ser la situación más frecuente. Como ejemplo vamos a ver los valores de PTH al año en función de los estadios de IRC, utilizando el estadio "sin IRC" como categoría de referencia. Esta variable toma los valores 1, 3, 4 y 5. Para realizar los contrastes y las comparaciones, la categoría 1 (de referencia) debe tomar el valor 0, y las demás categorías tomar el valor 1 en cada contraste. La creación de las distintas variables tendría la sintaxis siguiente:

```
COMPUTE IRC3 = (EstadIRC=3).
COMPUTE IRC4 = (EstadIRC=4).
COMPUTE IRC5 = (EstadIRC=5).
EXECTUE.
```

De esta manera cuando el estadio de IRC es 3, 4 o 5 pasa a tomar el valor 1, mientras que cuando la categorías es sin IRC (estadio 1) tomará el valor 0 y será la de referencia.

Posteriormente realizamos el procedimiento de regresión lineal introduciendo estas nuevas variables creadas:

```
REGRESSION
/MISSING LISTWISE
```

```
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT PTH4
/METHOD=ENTER IRC3 IRC4 IRC5.
```

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	97235,694	2	48617,847	4,254	,019[b]
	Residuo	605657,546	53	11427,501		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), IRC4, IRC3

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	107,028	21,380		5,006	,000	64,145	149,911
	IRC3	67,331	29,671	,300	2,269	,027	7,820	126,843
	IRC4	136,622	57,567	,314	2,373	,021	21,157	252,087

a. Variable dependiente: PTH4

En la tabla anterior vemos los resultados tras la ejecución de la sintaxis. Vemos que sólo aparece IRC3 e IRC4 porque no hay ningún paciente con estadio 5 de IRC.

Los coeficientes anteriores significan lo siguiente: los pacientes con estadio 3 de IRC (estadio 3=IRC3) tiene 67.33 pg/ml más de PTH que los pacientes sin IRC, mientras que los pacientes con estadio 4 (estadio 4=IRC4) tienen 136.62 pg/ml más de PTH que los pacientes sin IRC; ambos son estadísticamente significativos, con p= 0.027 para el primero y p=0.021 para el segundo. Hay que anotar que esta comparación sólo tiene sentido cuando se introducen todos las subcategorías creadas juntas en el modelo de regresión, no sirven por separado.

Si realizamos un ANOVA con estas mismas variables y los mismos contrastes obtendremos los mismos resultados y la misma significación, como vemos a continuación:

Coeficientes de contraste

	EstadIRC		
Contraste	Sin IRC	Estadio 3	Estadio 4
1	-1	1	0
2	-1	0	1

ANOVA

PTH4

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	97235,694	2	48617,847	4,254	,019
Dentro de grupos	605657,546	53	11427,501		
Total	702893,240	55			

Pruebas de contraste

		Contraste	Valor de contraste	Error estándar	t	gl	Sig. (bilateral)
PTH4	Suponer varianzas iguales	1	67,331	29,6705	2,269	53	,027
		2	136,622	57,5671	2,373	53	,021
	No se asume varianzas iguales	1	67,331	28,5191	2,361	33,179	,024
		2	136,622	62,0597	2,201	3,164	,110

Una vez visto el modelo de regresión con una sola variable independiente, vamos a ver cómo se interpretan los coeficientes cuando se introducen más de una variable, que sería el "Análisis Multivariante". Para ello vamos a utilizar como ejemplo la valoración de los niveles de PTH al año en función de la presencia de IRC y de los valores de PTH basales al momento del trasplante. Tras introducir ambas variables en el modelo de regresión y tras ejecutar la sintaxis obtenemos los siguientes resultados:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT PTH4
/METHOD=ENTER IRC_CKD PTHPrevTx.
```

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,547[a]	,299	,273	96,3974

a. Predictores: (Constante), PTHPrevTx, IRC_CKD

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	210392,617	2	105196,308	11,321	,000[b]
	Residuo	492500,623	53	9292,465		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), PTHPrevTx, IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	60,142	22,999		2,615	,012	14,011	106,273
	IRC_CKD	62,790	26,162	,279	2,400	,020	10,316	115,265
	PTHPrevTx	,136	,036	,434	3,739	,000	,063	,210

a. Variable dependiente: PTH4

En primer lugar vemos que al introducir otra variable en el modelo de regresión ha aumentado la parte explicada por la recta sobre la variabilidad total de la PTH al año; en efecto, el valor del R2 ha pasado de 0.115 a 0.299 (ha pasado de explicar el 11.5% al 29.9%). Por tanto podemos sacar como conclusión de lo anterior, que a mayor número de variables predictoras, mayor capacidad explicativa del modelo. Esto lo tendremos en cuenta más adelante cuando hablemos de la variante predictiva de los modelos de regresión. Pero este R2 aumenta sólo por el hecho de introducir más variables aunque algunas de ellas no tengan capacidad predictiva, por eso en el análisis multivariante se utiliza el coeficiente de determinación ajustado ("R cuadrado ajustado") que en este caso vale 0.273.

El valor de la nueva ecuación sería: PTH= 60.1 + 62.8 x IRC-CKD + 0.14 x PTHPrevTx.

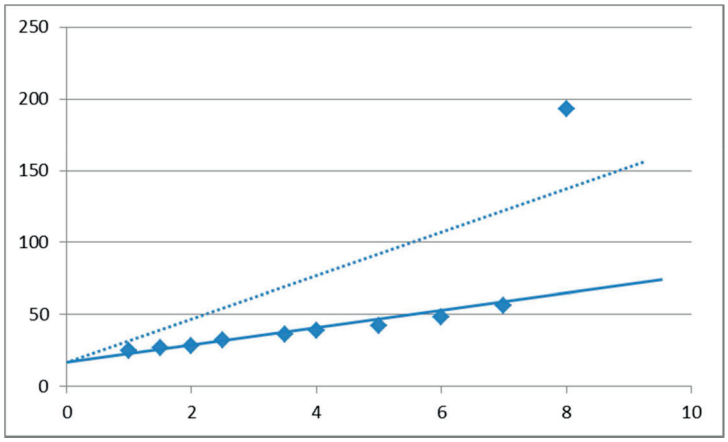
El modelo sigue siendo estadísticamente significativo, con una p<0.001. ¿Cómo se interpretan los coeficientes B de cada variable? El coeficiente de cada variable representa el valor de esa variable una vez extraído el efecto de la otra. Si observamos la ecuación de la recta, el coeficiente de cada variable nos está indicando los cambios que se producen en la PTH al año cuando la otra variable se mantiene constante (sin cambios). Vemos que el coeficiente de IRC-CKD ha pasado de valer 76.3 en el análisis univariante a 62.8 en el multivariante. Es decir, los pacientes con IRC (categoría 1) tras ajustar por la PTH basal, tienen 62.8 pg/ml más de PTH que los pacientes sin IRC. Por tanto es la influencia de una variable ajustada por las demás. En este caso también podríamos decir que: ajustado por la presencia de IRC, por cada pg/ml de aumento de la PTH basal, se produce un aumento de 0.14 pg/ml de la PTH al año.

¿Cuál de las 2 variables aporta más a la variación de la PTH al año? Para ello nos fijamos en los valores de los "coeficientes estandarizados Beta" de cada variable. En este caso vemos que la PTH basal aporta más que la presencia de IRC porque tiene un coeficiente Beta de 0.434.

Diagnósticos del modelo de regresión lineal

Una vez establecido el modelo de regresión, ¿cómo podemos saber si el modelo con las variables seleccionadas es el adecuado? Al comienzo del tema vimos que la mejor recta de regresión era la que tenía los residuales más pequeños. Por tanto la valoración del modelo se realizará analizando los residuales que representan la parte no explicada por la recta de regresión.

El problema surge cuando existen muchos casos en la muestra recogida con valores anómalos en alguna variable (tanto dependiente como independiente) porque son los que van a dar lugar a residuales mayores y a que el modelo estimado no ajuste correctamente a la mayoría de los datos.



Si observamos el gráfico anterior podemos ver que la recta de regresión representada con puntos viene determinada por la presencia de un valor extremadamente elevado en la variable Y del caso 8. Visualmente podemos observar que esta recta no es la que representa correctamente al

resto de los datos, sino que visualmente "ajusta" mejor la recta representada como una línea continua. ¿Qué hacer en este caso? En primer lugar debemos revisar los datos y comprobar si el valor recogido es correcto o no. Si es correcto, una medida sería eliminarlo de la base de datos y estimar nuevamente la recta, pero realmente estaríamos eliminando valores plausibles y la nueva recta no serviría para valores tan alejados. Queda a juicio del investigador qué hacer con estos valores.

Generalmente estos problemas surgen cuando el tamaño de muestra es reducido y obtenemos casos con valores extremos. Estos valores al aumentar el tamaño de la muestra se repetirían en más casos y quedarían "camuflados".

Por otra parte cuantas más variables introduzcamos en el modelo de regresión, más casos necesitamos de cada combinación de variables para que la recta de regresión se estime de manera adecuada. En el anterior análisis multivariante es fácil obtener casos con distintos valores de PTH basal en pacientes con IRC y sin IRC, pero si introducimos otra variable como el Sexo debemos tener suficientes casos registrados en cada combinación (con y sin IRC, en hombres y mujeres, y con distintas PTH basales) para estimar el modelo. Por eso se suele requerir que el tamaño de la muestra sea al menos

10 veces el número de variables del modelo de regresión. Así, si introducimos 3 variables en el modelo de regresión necesitamos al menos 30 casos para que se puedan dar todas las combinaciones posibles con un número suficiente de casos en cada una de ellas.

Para el análisis diagnóstico del modelo de regresión y asegurarnos que éste es adecuado, vamos a utilizar los siguientes índices:

- Residual externamente estandarizado.
- Valor de influencia centrado.
- Distancia de Cook.
- Tolerancia.
- Comprobación de normalidad de los residuales.

Todos ellos se obtienen a través del SPSS. Para ello vamos a coger el ejemplo anterior para ilustrar este apartado.

Residual Externamente Estandarizado (SDRESID)

Este índice detecta la presencia de valores alejados de la variable Y. Un valor de SDRESID superior a 2, y especialmente superior a 3 en algún caso, nos debe replantear la posibilidad de eliminar ese caso de la base de datos.

Valor de influencia centrado (LEVER)

Detecta la presencia de valores alejados de la variable X. Un valor de LEVER superior a $2(p+1)/n$, donde "p" es el número de variables del modelo estimado y "n" el tamaño de la muestra, nos debe replantear la posibilidad de eliminar ese caso de la base de datos.

Distancia de Cook (COOK)

Este índice detecta la presencia de valores alejados tanto de la variable X como Y que influyen notablemente en la estimación de los parámetros del modelo seleccionado. Un valor de COOK superior a 1 nos debe replantear la posibilidad de eliminar ese caso de la base de datos.

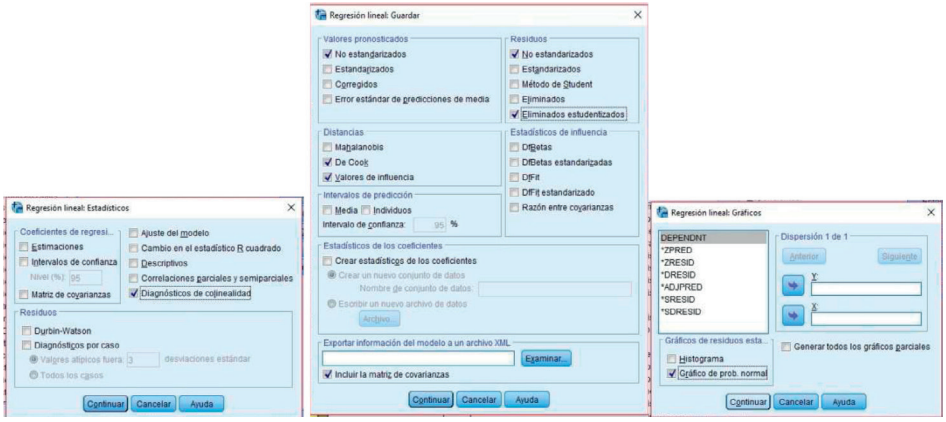
Tolerancia (TOL COLL)

Nos marca la presencia de colinealidad entre variables. Un valor de Tolerancia inferior a 0.1 nos está indicando que hay variables que están aportando la misma información (información redundante) y por tanto habría que eliminar del modelo alguna de ellas.

Análisis de la Normalidad de los residuales

Si el modelo estimado se ajusta bien a los datos, los residuales seguirán una distribución normal de media 0. Se evalúa mediante Shapiro-Wils o Kolmogorov-Smirnov.

Vamos a ilustrar los métodos diagnósticos con el ejemplo anterior donde la variable dependiente era "PTH al año" y las variables independientes "PTH basal" e "IRC". Para obtener los índices anteriores, en el cuadro de diálogo de la regresión debemos hacer Clic en el botón "Estadístico" y seleccionar la opción "Diagnóstico de colinealidad" (TOL COLL). Posteriormente hacemos Clic en el botón "Guardar" y seleccionamos las opciones "Valores pronosticados no Estandarizados", "Distancias De Cook" (COOK), "Valores de influencia" (LEVER), "Residuales No estandarizados" y "Residuales Eliminados estudentizados" (SDRESID). Por último hacemos Clic en el botón "Gráficos" y seleccionamos la opción "Gráfico de Prob. Normal". Pegamos la sintaxis porque posteriormente hay que editarla "a mano". Los cuadros de diálogos y la sintaxis obtenida son los siguientes:




```
 REGRESSION
/MISSING LISTWISE
/STATISTICS COLLIN TOL
```

/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT PTH4
/METHOD=ENTER PTHPrevTx IRC_CKD
/RESIDUALS NORMPROB(ZRESID)
/SAVE PRED COOK LEVER RESID SDRESID

Esta sintaxis hay que editarla: En "/RESIDUALS" quitar NORMPROB(ZRESID) y poner OUTLIERS (SDRESIS LEVER COOK); de esta manera nos listará los 10 casos con los valores de estos 3 índices más elevados. La sintaxis quedaría así:

REGRESSION
/MISSING LISTWISE
/STATISTICS COLLIN TOL
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT PTH4
/METHOD=ENTER PTHPrevTx IRC_CKD
/RESIDUALS OUTLIERS (SDRESIS LEVER COOK)
/SAVE PRED COOK LEVER RESID SDRESID.

Tras ejecutarla tenemos los siguientes resultados:

Coeficientes[a]

Modelo		Estadísticas de colinealidad	
		Tolerancia	VIF
1	PTHPrevTx	,981	1,019
	IRC_CKD	,981	1,019

a. Variable dependiente: PTH4

Estadísticas de residuos[a]

	Mínimo	Máximo	Media	Desviación estándar	N
Valor pronosticado	67,919	343,148	149,250	61,8492	56
Valor pronosticado estándar	-1,315	3,135	,000	1,000	56
Error estándar de valor pronosticado	17,339	52,816	21,525	5,926	56
Valor pronosticado corregido	68,384	365,762	150,138	64,1218	56
Residuo	-129,0615	314,6523	,0000	94,6285	56
Residuo estándar	-1,339	3,264	,000	,982	56
Residuo estudentizado	-1,382	3,321	-,004	1,009	56
Residuo eliminado	-137,4230	325,7197	-,8876	100,1656	56
Residuo estudentizado suprimido	-1,394	3,697	,010	1,054	56
Distancia de Mahal.	,797	15,529	1,964	2,360	56
Distancia de Cook	,000	,261	,020	,042	56
Valor de influencia centrado	,014	,282	,036	,043	56

a. Variable dependiente: PTH4

En la primera tabla podemos ver la colinealidad mediante el valor de la Tolerancia. Vemos que los valores son superiores a 0.1 y por tanto no existe colinealidad entre ellas. En la segunda tabla vemos los valores de SDRESID (Residuo estudentizado suprimido) y vemos que el valor máximo es de 3.697 (superior a 3) y por tanto habrá algún caso de la variable Y (PTH al año) con valor alejado. La distancia de Cook es inferior a 1 y por tanto no habrá ningún caso con valores que influyan notablemente en la estimación de los parámetros. En Valor de

influencia centrado (que en este caso debe ser inferior a 0.107) vemos que el valor máximo es de 0.282 (superior a 0.107) y por tanto habrá algún caso con valores en la variable X (combinación de PTH basal e IRC) con valores alejados. Al ejecutar la sintaxis, además de los resultados anteriores, se crean 5 nuevas variables con los valores de los índices anteriores en cada caso. Además se obtiene la siguiente tabla donde se muestran los 10 casos con los valores SDRESIS, COOK y LEVER más altos.

Estadísticas de valor atípico[a]

		Número del caso	Estadístico	Sig. F
Residuo estudentizado suprimido	1	30	**3,697**	
	2	17	**3,095**	
	3	10	**2,801**	
	4	36	1,849	
	5	22	-1,394	
	6	21	-1,297	
	7	6	1,249	
	8	23	-1,198	
	9	24	-1,186	
	10	5	1,079	
Distancia de Cook	1	17	,261	,853
	2	30	,129	,942
	3	51	,098	,961
	4	10	,092	,964
	5	2	,080	,971
	6	22	,041	,989
	7	36	,037	,991
	8	31	,025	,995
	9	43	,023	,995
	10	21	,023	,995
Valor de influencia centrado	1	51	**,282**	
	2	2	**,211**	
	3	17	,069	
	4	43	,057	
	5	9	,051	
	6	31	,047	
	7	29	,044	
	8	22	,043	
	9	37	,041	
	10	4	,038	

a. Variable dependiente: PTH4

En negrita están marcados los valores de SDRESID con valores superiores a 2 y de LEVER con valores superiores a 0.107.

Nos queda por analizar la Normalidad de los residuales. Para ello valoramos mediante las pruebas habituales sobre los residuales de la variable guardada en la base de datos RES_1. El resultado es el siguiente:

Pruebas de normalidad

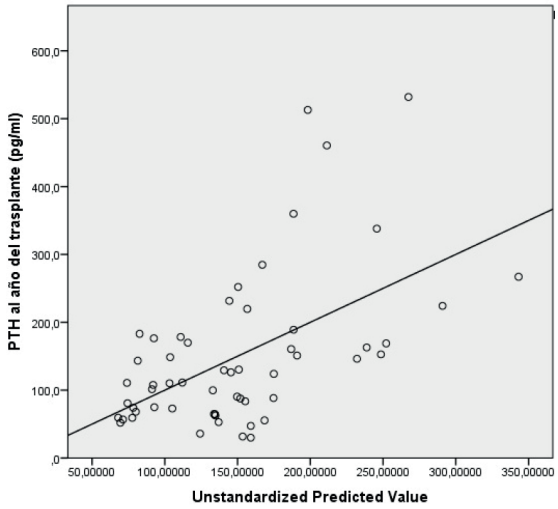
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
RES_1	,153	56	,002	,886	56	,000

a. Corrección de significación de Lilliefors

Vemos que el valor es significativo en cualquiera de las 2 pruebas, por tanto los residuales no siguen una distribución normal y el modelo estimado no sería adecuado.

Los estudios diagnósticos anteriores tienen su mayor interés generalmente cuando la regresión se utiliza con fines diagnósticos o predictivos, es decir, cuando vamos a seleccionar una serie de variables que nos expliquen los resultados encontrados en la variable independiente, o cuando vamos a querer predecir cuál sería el valor de la variable dependiente. En estos 2 casos necesitamos asegurarnos que las variables del modelo elegido son las correctas (aportan información relevante y sustancial sobre la variable independiente) y que los valores pronosticados con el modelo tengan el menor error posible. El problema surge cuando encontramos casos con valores anómalos en los índices diagnósticos anteriores. Qué hacemos, ¿los eliminamos de la base de datos? Esto tiene 2 problemas: el primero es que reducimos el tamaño de la muestra, y estos problemas generalmente suceden precisamente porque el tamaño de muestra es pequeño. Por otra parte, como veremos más adelante, hasta llegar al modelo final seleccionado, se han desechado por el camino otras variables que a lo mejor hubieran entrado en el modelo final si no existieran esos valores anómalos; por tanto requeriría volver a empezar de nuevo con el proceso de selección de variables. Como vemos no hay una respuesta contundente a la pregunta anterior. Si se detectan valores anómalos, en primer lugar comprobar que son correctos y no errores de la recogida de datos. Si no hay errores y los valores son plausibles, una forma de determinar si eliminarlos o no es realizando un análisis visual de la nube de puntos resultante de enfrentar los valores de la variable Y (variable respuesta) y los valores predichos por el modelo de regresión (que aparecen en la tabla de la base de datos como PRE_1). Si visualmente vemos que hay valores anómalos, pero que no están afectando significativamente a los resultados del modelo, los dejamos. Si vemos que se separan demasiado de la mayoría de los puntos, los eliminamos. Por otra parte, dado que el índice de COOK nos informa sobre problemas en la estimación de los parámetros del modelo, podemos darle más importancia a lo detectado con este índice frente a los demás, sobre todo si el objetivo de la regresión es con carácter predictivo.

El gráfico resultante de comparar la variable dependiente con los valores pronosticados del ejemplo utilizado es el siguiente:



Si los valores pronosticados por el modelo coinciden con los de la variable Y, todos los puntos se encontrarán sobre la recta. Vemos visualmente que a medida que aumentan los valores de PTH pronosticados por el modelo se separan cada vez más de la recta; por tanto, este modelo no sería muy válido para pronosticar valores de PTH elevados (vemos que para valores pronosticados de PTH de 200, hay pacientes con 150 y hasta 500).

Por tanto el modelo anterior no sería adecuado (recordar que el diseño de la base de datos del ejemplo no es con carácter descriptico o predictivo).

Modelo de regresión lineal para medir un efecto

Esta aplicación de la regresión lineal se utiliza para evaluar la influencia de una exposición (variable predictora) en la aparición de un evento, ajustado por otras variables que se han demostrado previamente que también influyen en la aparición del evento, denominadas "variables de confusión". Es por tanto una aplicación de la regresión indicada para estudios de seguimiento, donde los pacientes se van a dividir según presenten o no la exposición y se seguirán hasta la aparición del evento o hasta el fin del estudio. Será por tanto aplicable a estudios de Cohortes o Experimentales (Ensayos Clínicos).

Al seleccionar los pacientes en función de si presentan o no la exposición, si la selección no ha sido aleatorizada y con un tamaño grande de muestra, es bastante probable que los grupos no estén balanceados en determinadas variables que se hayan demostrado previamente que también influyen en la aparición del evento y por tanto no vamos a poder saber si la variable exposición a estudio es la responsable de la aparición del evento o es debido a esas otras variables. En los Ensayos Clínicos, donde suele haber un tamaño de muestra previo determinado, y donde la selección de los pacientes a uno u otro grupo se hace de manera aleatoria (randomizado), este problema se reduce, pero aun así, hay que comprobar antes del análisis si existe o no disbalances entre los grupos en determinadas variables, y de haberlos debemos ajustar por ellas.

Como hemos comentado anteriormente, el objetivo es ajustar por aquellas variables que se hayan demostrado previamente en otros estudios que también influyen en la aparición del evento. Por tanto, previo a la realización de cualquier estudio, es obligado el haber realizado una exhaustiva revisión bibliográfica para conocer cuáles son las variables que pueden influir en los resultados y debemos también recogerlas en nuestra base de datos.

Para ilustrar esta aplicación de la regresión lineal, vamos a utilizar el ejemplo de nuestra base de datos. Aunque no es el objetivo del estudio original, vamos a ver la influencia de la IRC sobre los valores de PTH al año. En otros estudios se ha demostrado que los valores de PTH al año del trasplante dependen de los valores de PTH al momento del trasplante, del tiempo previo en diálisis y del tipo de diálisis previa. Por tanto, queremos ver la influencia de la IRC sobre los valores de PTH ajustada por el resto de variables.

Teóricamente, el ajuste consistiría simplemente en introducir en la regresión todas las variables registradas y ya está. El efecto de la variable exposición sería el ajustado al resto de variables. El problema aparece cuando tenemos un gran número de variables de confusión o cuando hemos recogido variables cuya influencia sobre la aparición del evento no está clara-

mente demostrada en otros estudios. ¿Debemos ajustar por todas las variables? Si hemos registrado muchas variables y ajustamos por todas, con un tamaño de muestra no muy grande, vamos a tener problemas a la hora de estimar el efecto de la variable de exposición, con grandes fluctuaciones en su valor según la variable de confusión introducida y con intervalos de confianza excesivamente grandes. Si hemos registrado variables que no están claramente demostrado su influencia, primero deberemos ver si en nuestra muestra realmente es o no una variable de confusión.

Vamos a considerar que una variable es variable de confusión si cumple fundamentalmente 2 condiciones:

1) se ha demostrado claramente en otros estudios que influye de manera significativa en la aparición del evento que estamos estudiando, y por tanto deberemos ver si esta variable está disbalanceada en los grupos de nuestra variable de exposición. Consistiría en ver por tanto si hay o no diferencias significativas en las distintas variables en los grupos de exposición. En nuestro ejemplo consistiría en ver si hay diferencias significativas de las variables PTH al momento del trasplante, tiempo en diálisis y tipo de diálisis previa en los grupos con y sin IRC. Esta comparación se realizará mediante análisis univariantes mediante t-Student, Chi-cuadrado, etc., según corresponda. Se considera variable de confusión cuando la significación es estadística ($p<0.05$). Si la "p" es mayor, se considerará igualmente como confusión si nosotros lo estimamos oportuno según los conocimientos teóricos.

2) la variable en cuestión es un factor de riesgo o protector de la aparición del evento en el grupo no expuesto. En nuestro ejemplo consistiría en comprobar en el grupo sin IRC si estas variables influyen en los niveles de PTH. Para ello debemos seleccionar sólo los pacientes sin IRC y realizar análisis de regresión lineal univariantes con estas variables. Generalmente se consideran variables de confusión cuando la significación del modelo de regresión presenta una $p<0.2$.

Para que sean variables de confusión deben cumplir los dos requisitos anteriores. Pero si el investigador cree que no introducir una determinada variable como ajuste supone un importante sesgo debido a lo demostrado en otros estudios, se debe introducir aunque no cumpla los 2 criterios anteriores. Esto se planteará generalmente cuando cumpla un criterio y no otro, sobre todo cuando cumpla el primer criterio y no el segundo, o cuando cumpla el segundo criterio y el primero se quede al borde de la significación estadística.

Vamos a ver cómo se realizaría con SPSS con los datos de nuestro ejemplo:

Primer criterio

Debemos ver si existe disbalance en las 3 variables de confusión en los grupos con y sin IRC. Para ello vamos a realizar comparación de medias en muestras independientes para las variables cuantitativas (PTH al momento del trasplante y Tiempo en diálisis) y tablas de contingencia para las categóricas (Tipo de diálisis previa). Valoraremos previamente como es lógico si las variables cuantitativas siguen o no distribución Normal en las 2 categorías, y si la frecuencia esperada es inferior a 5 en alguna casilla de la tabla de contingencia. Si la varia-

ble a valorar el efecto fuese cuantitativa (por ejemplo fuese el Filtrado Glomerular en ml/min en lugar de la presencia o no de IRC) el análisis se haría a través de análisis de regresión lineal univariantes con cada una de las variables a valorar la confusión.

En este caso, tanto el Tiempo en diálisis como la PTH al momento del trasplante no siguen distribución normal y por tanto usaremos para valorar la significación estadística pruebas no paramétricas. Los resultados son los siguientes:

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de PTH previa al trasplante (pg/ml) es la misma entre las categorías de IRC mediante CKD-EPI.	Prueba U de Mann-Whitney para muestras independientes	,212	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de Tiempo en Diálisis previo al trasplante (meses) es la misma entre las categorías de IRC mediante CKD-EPI.	Prueba U de Mann-Whitney para muestras independientes	,212	Conserve la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significación es ,05.

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	,723[a]	1	,395		
Corrección de continuidad[b]	,277	1	,599		
Razón de verosimilitud	,738	1	,390		
Prueba exacta de Fisher				,522	,302
Asociación lineal por lineal	,711	1	,399		
N de casos válidos	61				

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 5,31.

b. Sólo se ha calculado para una tabla 2x2

Vemos que ninguna de las tres variables son estadísticamente significativas, es decir, no están disbalanceados los dos grupos en función de estas variables. Por tanto, no serían consideradas como variables de ajuste y no haría faltar comprobar el segundo criterio. No obstante vamos a hacerlo para ver cómo se hace.

Segundo criterio

En primer lugar debemos seleccionar de la base de datos sólo los pacientes sin IRC. Para ello debemos realizar el siguiente paso previo al análisis:

Datos → Seleccionar casos → Si se satisface la condición → Si… → IRC=0; Cuya sintaxis es:

USE ALL. COMPUTE filter_$=(IRC_CKD = 0). VARIABLE LABELS filter_$ 'IRC_CKD = 0 (FILTER)'. VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0).

FILTER BY filter_$.

EXECUTE.

A continuación realizamos análisis de regresión lineal univariante con las tres variables a analizar. Los resultados serían:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	86,880	13,330		6,518	,000
	PTHPrevTx	,059	,028	,406	2,129	,044

a. Variable dependiente: PTH4

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	106,133	11,196		9,480	,000
	TipoDiaPrev	5,592	27,989	,042	,200	,843

a. Variable dependiente: PTH4

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	97,738	20,312		4,812	,000
	TiempHD	,169	,319	,110	,529	,602

a. Variable dependiente: PTH4

En este caso vemos que tan sólo la variable "PTH previa la trasplante" tiene una $p<0.2$, por tanto esta variable podría ser una variable de confusión (si cumpliese también el primer criterio).

Vamos a suponer que la "PTH previa al trasplante" hubiese cumplido los 2 criterios. Introduciríamos en el modelo de regresión la variable "IRC-CKD" que es la variable exposición y la variable "PTH previa al trasplante" que sería la variable de confusión.

NOTA: ¡tras realizar un proceso de selección de casos mediante el procedimiento anterior, no olvidar posteriormente volver a seleccionar todos los casos!

El resultado tras el ajuste sería:

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	210392,617	2	105196,308	11,321	,000[b]
	Residuo	492500,623	53	9292,465		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), PTHPrevTx, IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	60,142	22,999		2,615	,012	14,011	106,273
	IRC_CKD	62,790	26,162	,279	2,400	,020	10,316	115,265
	PTHPrevTx	,136	,036	,434	3,739	,000	,063	,210

a. Variable dependiente: PTH4

Diríamos por tanto, que tras ajustar por los valores de PTH previos al trasplante, la presencia de IRC aumenta los niveles de PTH al año una media de 62.79 pg/ml (IC 95%: 10.32 a 115.27, p=0.02).

Modelo de regresión lineal con finalidad descriptiva

En este caso la utilidad de la regresión es muy diferente a la del apartado anterior. Se trata de ver qué variables pueden determinar la aparición de un evento. Esta finalidad de la regresión, junto con la finalidad predictiva, son las más utilizadas. Su utilidad se centra en la valoración de las posibles causas que puedan dar lugar a un evento observado, característico de los estudios retrospectivos (Casos-Controles). En este tipo de estudio los pacientes se eligen según presenten o no el evento tratando de valorar cuáles son las características previas de los pacientes que motivan la aparición de ese evento mediante una revisión de los antecedentes de los pacientes en las historias clínicas. Suelen ser pasos previos a la realización de los estudios prospectivos como Ensayos Clínicos y Cohortes. Por ejemplo, imaginemos que en nuestra práctica clínica diaria hemos observado en la consulta que algunos pacientes presentan niveles de ácido úrico más elevados que otros y queremos ver cuáles son las causas que dan lugar a esta diferencia. Revisaremos las historias clínicas de los pacientes recogiendo determinadas variables para ver cuáles son las responsables del evento observado. Son estudios exploratorios en busca de posibles causas para posteriormente realizar estudios prospectivos.

Una de las características de esta aplicación de la regresión es conseguir modelos lo más simples posibles, de tal forma que con la mínima cantidad de variables posibles se consiga dar la mayor información posible. Es un principio de parsimonia. De todas las posibles variables registradas para un estudio escogeremos sólo aquellas más relevantes siguiendo este principio. Como vimos en anteriores apartados, la cantidad de variabilidad que es capaz de explicar el modelo de regresión venía determinada por el coeficiente de determinación R2, y por tanto las variables que formen parte del modelo de regresión serán aquellas que tengan el mayor R2. Podemos intuir que podemos tener varios modelos de regresión con diferentes variables que tengan un mismo R2 o parecido. ¿Cuál escogemos entonces? Escogeremos aquél que tenga el menor número de variables posibles (parsimonia) o aquél cuyas variables sean más idóneas según criterios del investigador teniendo en cuenta otras consideraciones (como coherencia biológica, facilidad para su registro sin errores, etc.).

Vamos a ver cómo seleccionamos las variables que van a formar parte del mejor modelo de regresión mediante SPSS. Aunque la base de datos no está diseñada para tal fin (está diseñada para un estudio prospectivo y valorar un efecto) vamos a utilizarla de ejemplo. En este caso vamos a imaginar que seleccionamos a los pacientes según los valores de PTH al año del

trasplante y queremos ver qué variables de las registradas son las que determinan las diferencias que encontramos entre unos pacientes y otros.

El programa SPSS tiene instalados tres procedimientos automáticos para la selección de variables: una inclusión secuencial (Forward selection), una exclusión secuencial (Backward elimination), y una selección por pasos (Stepwise regression).

Inclusión secuencial "Forward selection"

En este procedimiento el programa va introduciendo variables al modelo de regresión en función del grado de significación de cada variable, terminando el procedimiento cuando las variables restantes no cumplen los criterios de inclusión, generalmente cuando el valor p es mayor de 0.05. Cuando introduce una variable el programa analiza la correlación del resto de variables con la variable dependiente e introduce en el modelo aquella que sea más significativa. Es decir, primero introduce la que tenga el menor valor "p", analiza las restantes, introduce la siguiente con menor valor "p", analiza la restantes, introduce la siguiente con menor "p", y así sucesivamente hasta que las variables restantes que aún no se han introducido en el modelo tengan un valor superior a 0.05 que suele ser el predeterminado. Vamos a ver cómo se haría con SPSS imaginando que se han registrado las variables Sexo, Edad, Tipo de diálisis previa, PTH al momento del trasplante, Tiempo en diálisis y presencia de IRC al año mediante CKD- EPI. El procedimiento con SPSS sería el siguiente:

Analizar → Regresión → Lineales; seleccionamos como variable dependiente "PTH al año". En el cuadro de variables independientes seleccionamos las variables a analizar. En la opción "Método" seleccionamos "hacia adelante". En el botón "Estadísticos" seleccionamos la opción "Intervalos de confianza 95%". La sintaxis completa y el cuadro de diálogo serían los siguientes:


```
REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10) /NOORIGIN
/DEPENDENT PTH4
/METHOD=FORWARD Sexo Edad TipoDiaPrev PTHPrevTx IRC_CKD TiempHD.
```

Tras ejecutar la sintaxis obtenemos lo siguiente:

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	PTHPrevTx	.	Avanzar (Criterio: Probabilidad-de-F-para-entrar <= ,050)
2	IRC_CKD	.	Avanzar (Criterio: Probabilidad-de-F-para-entrar <= ,050)

a. Variable dependiente: PTH4

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,472[a]	,223	,209	100,5565
2	,547[b]	,299	,273	96,3974

a. Predictores: (Constante), PTHPrevTx

b. Predictores: (Constante), PTHPrevTx, IRC_CKD

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	156865,919	1	156865,919	15,513	,000[b]
	Residuo	546027,321	54	10111,617		
	Total	702893,240	55			
2	Regresión	210392,617	2	105196,308	11,321	,000[c]
	Residuo	492500,623	53	9292,465		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), PTHPrevTx

c. Predictores: (Constante), PTHPrevTx, IRC_CKD

Variables excluidas[a]

Modelo		En beta	t	Sig.	Correlación parcial	Estadísticas de colinealidad
						Tolerancia
1	Sexo	,066[b]	,548	,586	,075	1,000
	Edad	,013[b]	,107	,915	,015	,992
	TipoDiaPrev	-,014[b]	-,113	,911	-,015	1,000
	IRC_CKD	,279[b]	2,400	,020	,313	,981
	TiempHD	,078[b]	,627	,533	,086	,935
2	Sexo	-,017[c]	-,140	,889	-,019	,912
	Edad	-,070[c]	-,576	,567	-,080	,915
	TipoDiaPrev	-,047[c]	-,406	,686	-,056	,986
	TiempHD	,019[c]	,154	,878	,021	,893

a. Variable dependiente: PTH4

b. Predictores en el modelo: (Constante), PTHPrevTx

c. Predictores en el modelo: (Constante), PTHPrevTx, IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	90,092	20,153		4,470	,000	49,687	130,497
	PTHPrevTx	,149	,038	,472	3,939	,000	,073	,224
2	(Constante)	60,142	22,999		2,615	,012	14,011	106,273
	PTHPrevTx	,136	,036	,434	3,739	,000	,063	,210
	IRC_CKD	62,790	26,162	,279	2,400	,020	10,316	115,265

a. Variable dependiente: PTH4

En la primera tabla vemos las variables incluidas en el modelo de regresión final, en este caso las variables PTH previa al trasplante y presencia de IRC. Como podemos ver en la tercera tabla "Variables excluidas" en el Modelo 1 la siguiente variable cuya correlación es más significativa es la variable "IRC_CKD" introduciéndose la siguiente en el modelo. En el Modelo 2 vemos que ya no hay ninguna variable más que sea significativa, por eso ahí se

termina el proceso de selección. En la tabla "ANOVA" podemos ver que el modelo de regresión seleccionado es estadísticamente significativo con p<0.001 y con un R2 de 0.299 (Tabla 2 "Resumen del modelo"). Por último tenemos en la última tabla el valor de los coeficientes de cada variable seleccionada con sus intervalos de confianza. Según este modelo, los principales determinantes de los valores de PTH al año del trasplante son la PTH previa al trasplante y la presencia de IRC.

Exclusión secuencial "Backward elimination"

Mediante este procedimiento el programa hace lo contrario de lo comentado en el apartado anterior. El primer paso que realiza consiste en introducir en un modelo de regresión todas las variables predictoras y posteriormente ir eliminando una a una aquellas variables que sean menos significativas hasta que en el modelo ya no quede ninguna variable que cumpla el criterio de exclusión, generalmente prefijado en 0.1 (p>0.1). Este valor "p" prefijado se puede modificar "a mano" a través de la sintaxis situándolo en un valor superior a 0.1 si queremos que queden más variables al final en el modelo, o menor (por ejemplo 0.05) para que el proceso de eliminación sea más estricto y queden menos variables en el modelo final.

Vamos a ver cómo se realizaría mediante SPSS. El procedimiento sería el siguiente:

Analizar → Regresión → Lineales; seleccionamos como variable dependiente "PTH al año". En el cuadro de variables independientes seleccionamos las variables a analizar. En la opción "Método" seleccionamos "hacia atrás". En el botón "Estadísticos" seleccionamos la opción "Intervalos de confianza 95%". La sintaxis completa y el cuadro de diálogo serían los siguientes:

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN

/DEPENDENT PTH4
/METHOD=BACKWARD Sexo Edad TipoDiaPrev PTHPrevTx IRC_CKD TiempHD.

Tras ejecutar la sintaxis obtenemos lo siguiente:

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	TiempHD, TipoDiaPrev, Edad, PTHPrevTx, Sexo, IRC_CKD[b]	.	Entrar
2	.	TiempHD	Retroceder (criterio: Probabilidad de F-para-eliminar >= , 100).
3	.	Sexo	Retroceder (criterio: Probabilidad de F-para-eliminar >= , 100).
4	.	TipoDiaPrev	Retroceder (criterio: Probabilidad de F-para-eliminar >= , 100).
5	.	Edad	Retroceder (criterio: Probabilidad de F-para-eliminar >= , 100).

a. Variable dependiente: PTH4

b. Todas las variables solicitadas introducidas.

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	216587,389	6	36097,898	3,637	,005[b]
	Residuo	486305,851	49	9924,609		
	Total	702893,240	55			
2	Regresión	216575,738	5	43315,148	4,453	,002[c]
	Residuo	486317,502	50	9726,350		
	Total	702893,240	55			
3	Regresión	215789,897	4	53947,474	5,648	,001[d]
	Residuo	487103,343	51	9551,046		
	Total	702893,240	55			
4	Regresión	213514,883	3	71171,628	7,563	,000[e]
	Residuo	489378,357	52	9411,122		
	Total	702893,240	55			
5	Regresión	210392,617	2	105196,308	11,321	,000[f]
	Residuo	492500,623	53	9292,465		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), TiempHD, TipoDiaPrev, Edad, PTHPrevTx, Sexo, IRC_CKD

c. Predictores: (Constante), TipoDiaPrev, Edad, PTHPrevTx, Sexo, IRC_CKD

d. Predictores: (Constante), TipoDiaPrev, Edad, PTHPrevTx, IRC_CKD

e. Predictores: (Constante), Edad, PTHPrevTx, IRC_CKD

f. Predictores: (Constante), PTHPrevTx, IRC_CKD

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,555[a]	,308	,223	99,6223
2	,555[b]	,308	,239	98,6223
3	,554[c]	,307	,253	97,7295
4	,551[d]	,304	,264	97,0109
5	,547[e]	,299	,273	96,3974

a. Predictores: (Constante), TiempHD, TipoDiaPrev, Edad, PTHPrevTx, Sexo, IRC_CKD

b. Predictores: (Constante), TipoDiaPrev, Edad, PTHPrevTx, Sexo, IRC_CKD

c. Predictores: (Constante), TipoDiaPrev, Edad, PTHPrevTx, IRC_CKD

d. Predictores: (Constante), Edad, PTHPrevTx, IRC_CKD

e. Predictores: (Constante), PTHPrevTx, IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	105,931	71,521		1,481	,145	-37,796	249,657
	Sexo	-8,289	30,274	-,035	-,274	,785	-69,128	52,549
	Edad	-,809	1,227	-,088	-,659	,513	-3,274	1,657
	TipoDiaPrev	-16,012	33,457	-,059	-,479	,634	-83,247	51,222
	PTHPrevTx	,133	,039	,422	3,397	,001	,054	,211
	IRC_CKD	72,545	31,330	,322	2,315	,025	9,584	135,506
	TiempHD	-,007	,212	-,005	-,034	,973	-,433	,418
2	(Constante)	105,123	66,851		1,572	,122	-29,152	239,399
	Sexo	-8,435	29,674	-,036	-,284	,777	-68,036	51,167
	Edad	-,799	1,178	-,087	-,678	,501	-3,164	1,567
	TipoDiaPrev	-15,823	32,668	-,058	-,484	,630	-81,438	49,792
	PTHPrevTx	,132	,038	,421	3,504	,001	,056	,208
	IRC_CKD	72,286	30,100	,321	2,402	,020	11,829	132,743
3	(Constante)	99,239	62,990		1,575	,121	-27,219	225,697
	Edad	-,718	1,133	-,078	-,634	,529	-2,994	1,557
	TipoDiaPrev	-15,799	32,372	-,058	-,488	,628	-80,788	49,190
	PTHPrevTx	,133	,037	,423	3,563	,001	,058	,208
	IRC_CKD	69,303	27,956	,308	2,479	,017	13,179	125,426
4	(Constante)	92,713	61,102		1,517	,135	-29,897	215,323
	Edad	-,642	1,114	-,070	-,576	,567	-2,877	1,594
	PTHPrevTx	,134	,037	,425	3,609	,001	,059	,208
	IRC_CKD	67,197	27,418	,298	2,451	,018	12,179	122,214
5	(Constante)	60,142	22,999		2,615	,012	14,011	106,273
	PTHPrevTx	,136	,036	,434	3,739	,000	,063	,210
	IRC_CKD	62,790	26,162	,279	2,400	,020	10,316	115,265

a. Variable dependiente: PTH4

En la primera tabla vemos que en el primer paso se introducen todas las variables seleccionadas; en la segunda columna vemos qué variables se van eliminando en cada paso (la última variable eliminada es la Edad). El modelo seleccionado es estadísticamente significativo con una $p<0.001$ (tabla 2 ANOVA). En la tercera tabla "Resumen del modelo" vemos el valor de R2 que es de 0.299. En la última tabla vemos las variables seleccionadas para el modelo final. En este caso diremos que los principales determinantes de los valores de PTH al año del trasplante son la PTH previa al trasplante y la presencia de IRC.

Selección por pasos "Stepwise regression"

Este procedimiento es una mezcla de los dos anteriores. Se van introduciendo en el modelo los términos según la significación estadística. En un primer paso introduce el término con mayor significación estadística evaluando el resto de los términos introduciendo en el siguiente paso aquél que sea más significativo (como si fuese un procedimiento "hacia adelante"). Pero en este caso, una vez introducido el siguiente término, evalúa el modelo resultante y elimina del mismo aquella variable que cumpla con el criterio de exclusión (como si fuese un procedimiento "hacia atrás") a la vez que introduce la siguiente variable no introducida aún que sea más significativa. Es decir, va metiendo y sacando variables del modelo según cumpla con los criterios de inclusión (prefijados en $p<0.05$) o de exclusión ($p>0.1$). Vamos a ver cómo se realizaría con SPSS. El procedimiento es el siguiente:

Analizar → Regresión → Lineales; seleccionamos como variable dependiente "PTH al año". En el cuadro de variables independientes seleccionamos las variables a analizar. En la opción "Método" seleccionamos "Por pasos". En el botón "Estadísticos" seleccionamos la opción "Intervalos de confianza 95%". La sintaxis completa y el cuadro de diálogo serían los siguientes:


```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT PTH4
/METHOD=STEPWISE Sexo Edad TipoDiaPrev PTHPrevTx IRC_CKD TiempHD.
```

Tras ejecutar la sintaxis obtenemos lo siguiente:

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	PTHPrevTx	.	Por pasos (Criterios: Probabilidad-de-F-para-entrar <= ,050, Probabilidad-de-F-para-eliminar >= ,100).
2	IRC_CKD	.	Por pasos (Criterios: Probabilidad-de-F-para-entrar <= ,050, Probabilidad-de-F-para-eliminar >= ,100).

a. Variable dependiente: PTH4

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,472[a]	,223	,209	100,5565
2	,547[b]	,299	,273	96,3974

a. Predictores: (Constante), PTHPrevTx

b. Predictores: (Constante), PTHPrevTx, IRC_CKD

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	156865,919	1	156865,919	15,513	,000[b]
	Residuo	546027,321	54	10111,617		
	Total	702893,240	55			
2	Regresión	210392,617	2	105196,308	11,321	,000[c]
	Residuo	492500,623	53	9292,465		
	Total	702893,240	55			

a. Variable dependiente: PTH4

b. Predictores: (Constante), PTHPrevTx

c. Predictores: (Constante), PTHPrevTx, IRC_CKD

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	90,092	20,153		4,470	,000	49,687	130,497
	PTHPrevTx	,149	,038	,472	3,939	,000	,073	,224
2	(Constante)	60,142	22,999		2,615	,012	14,011	106,273
	PTHPrevTx	,136	,036	,434	3,739	,000	,063	,210
	IRC_CKD	62,790	26,162	,279	2,400	,020	10,316	115,265

a. Variable dependiente: PTH4

En la primera tabla vemos las variables que se introducen en cada paso. El modelo seleccionado es estadísticamente significativo con una $p<0.001$ (tabla 3 ANOVA). En la segunda tabla "Resumen del modelo" vemos el valor de R2 que es de 0.299. En la última tabla vemos las variables seleccionadas para el modelo final. En este caso diremos que los principales determinantes de los valores de PTH al año del trasplante son la PTH previa al trasplante y la presencia de IRC.

En este ejemplo hemos visto que los tres procedimientos seleccionan el mismo modelo de regresión final incluyendo las mismas variables. Esto generalmente no es lo que suele ocurrir seleccionando modelos diferentes (aunque con cierto parecido) cada procedimiento. ¿Con cuál nos quedamos? Como hemos comentado anteriormente, nos quedaremos con aquel que incluya menos variables (por el principio de parsimonia) sin que se produzcan grandes cambios en el R2. Es decir, si un modelo tiene 5 variables con un R2 de 0.365 y otro con 3 variables tiene un R2 de 0.342, nos quedaremos con éste último. Lógicamente al tener menos variables siempre va a tener un R2 más pequeño. Por otra parte, seremos nosotros mismos los que escojamos un modelo u otro según si las variables que forman parte de él "nos interesan o no" según los criterios que hemos comentado (plausibilidad biológica, rigor en la recogida de la variable, etc.).

En los anteriores ejemplos, el número de variables predictoras de entrada no era elevado (en esta caso eran 6 variables), pero es posible que en nuestros estudios hayamos recogida un número mayor de variables. En este caso se realiza un cribado previo de todas las variables mediante análisis de regresión univariantes seleccionando aquellas variables con un valor $p<0.2$ para la realización de los procedimientos anteriores. De esta manera evitaremos que ciertas variables puedan darnos problemas de colinealidad.

¿Cómo expresaríamos los resultados para una publicación? Realizaríamos una tabla como la siguiente donde en una primera parte daremos el valor de los coeficientes (IC95%, significación) del análisis univariante con cada variable, y en una segunda parte daremos el valor de los coeficientes (IC95%, significación) sólo de las variables del modelo que hayamos seleccionado.

	Univariante			Multivariante		
Variable	Coeficiente B	IC 95%	Significación	Coeficiente B	IC 95%	Significación
PTH previa						
…						

Por supuesto, tras la elección de un modelo de regresión, se deben comprobar la validez del mismo mediante las pruebas diagnósticas descritas en el apartado anterior.

Modelo de regresión lineal con finalidad predictiva

Esta aplicación de la regresión es similar a la anterior pero con una visión predictiva. Consiste en predecir el valor de la variable dependiente con la mayor exactitud posible, por lo que no se precisa del principio de simplicidad que habíamos comentado en el apartado anterior. La selección de las variables se realiza igual por alguno de los procedimientos anteriores, pero se puede ser menos estricto en la preselección inicial si tenemos muchas variables y podemos dar un valor "p" superior a 0.2.

En el ejemplo inicial del tema vimos que la ecuación de la recta de regresión seleccionada era:

PTH al año = 60.14 + 0.136*PTHprevia + 62.79*IRC_CKD

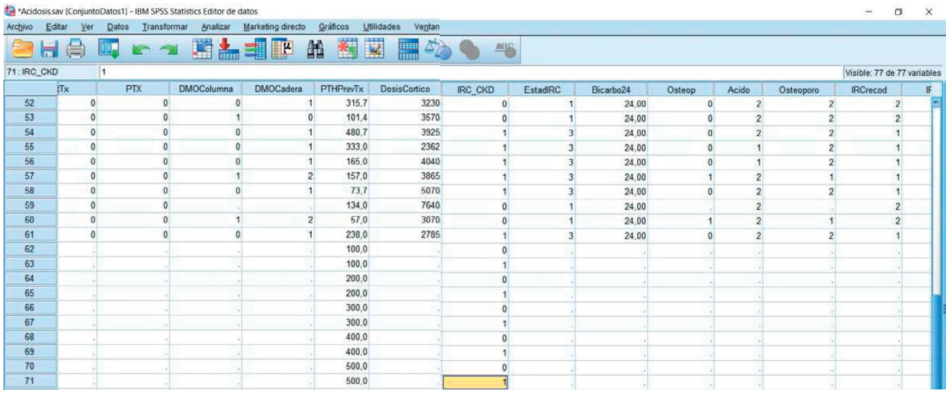
Conociendo los valores de PTH previa y de si presentan o no IRC podemos predecir el valor de PTH. Así para un paciente con PTH previa de 132 y con presencia de IRC, la PTH al año prevista sería de:

PTH al año = 60.14 + 0.136*132 + 62.79*1 = 140.88

De esta forma podemos predecir el valor de la variable dependiente para distintas combinaciones de valores de las variables predictoras. Este es el valor estimado, pero nos interesa también conocer cuál es el intervalo de confianza de esta estimación. Esta es la forma por la que se estiman las fórmulas utilizadas en la práctica clínica diaria como las fórmulas para estimar el filtrado glomerular (MDRD, CKD-EPI…), tablas de riesgos, etc.

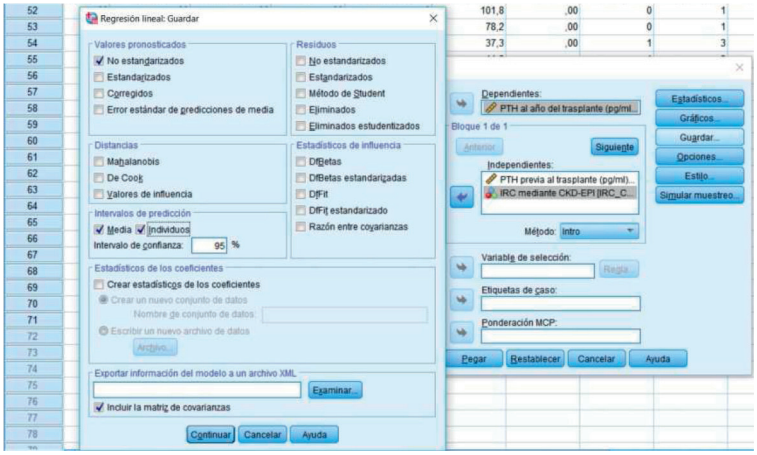
Para realizarlo con SPSS basta con introducir nuevos casos en la tabla de datos en las variables predictoras dejando vacías las casillas de la variable independiente y ejecutar posteriormente el modelo de regresión con las variables seleccionadas a través de un "Enter".

Vamos a introducir en la tabla de datos valores para la variable "PTHprev" de 100, 200, 300, 400, 500, con valores en IRC-CKD de 0 y 1 para cada valor; deberemos poner todas las opciones posibles tal como vemos en la imagen siguiente:



	fTx	PTX	DMOColumna	DMOCadera	PTHPrevTx	DosisCortico	IRC_CKD	EstadIRC	Bicarbo24	Osteop	Acido	Osteoporo	IRCrecod	IF
52	0	0	0	1	315,7	3230	0	1	24,00	0	2	2	2	
53	0	0	1	0	101,4	3570	0	1	24,00	0	2	2	2	
54	0	0	0	1	480,7	3925	1	3	24,00	0	2	2	1	
55	0	0	0	1	333,0	2362	1	3	24,00	0	1	2	1	
56	0	0	0	1	165,0	4040	1	3	24,00	0	1	2	1	
57	0	0	1	2	157,0	3865	1	3	24,00	1	2	1	1	
58	0	0	0	1	73,7	5070	1	3	24,00	0	2	2	1	
59	0	0	.	.	134,0	7640	0	1	24,00	.	2	.	2	
60	0	0	1	2	57,0	3070	0	1	24,00	1	2	1	2	
61	0	0	0	1	238,0	2785	1	3	24,00	0	2	2	1	
62	.	.	.	.	100,0	.	0	.	.	.	.	.	.	
63	.	.	.	.	100,0	.	1	.	.	.	.	.	.	
64	.	.	.	.	200,0	.	0	.	.	.	.	.	.	
65	.	.	.	.	200,0	.	1	.	.	.	.	.	.	
66	.	.	.	.	300,0	.	0	.	.	.	.	.	.	
67	.	.	.	.	300,0	.	1	.	.	.	.	.	.	
68	.	.	.	.	400,0	.	0	.	.	.	.	.	.	
69	.	.	.	.	400,0	.	1	.	.	.	.	.	.	
70	.	.	.	.	500,0	.	0	.	.	.	.	.	.	
71	.	.	.	.	500,0	.	1	.	.	.	.	.	.	

A continuación ejecutamos de nuevo la sintaxis con la variable dependiente "PTH al año" y como variables predictoras "PTHprevTx" y presencia de "IRC-CKD". En este caso en el cuadro de diálogo de la regresión debemos pedir en el botón "Guardar" que nos guarde los "Valores pronosticados No estandarizados" y los "Intervalos de predicción de la Media y de Individuos" como vemos en el siguiente cuadro de diálogo:



52		101,8	,00	0	1
53		78,2	,00	0	1
54		37,3	,00	1	3
55					
56					
57					
58					
59					
60					
61					
62					
63					
64					
65					
66					
67					
68					
69					
70					
71					
72					
73					
74					
75					
76					
77					
78					

Tras ejecutar la sintaxis veremos que se nos han creado 3 variables, una con los valores pronosticados (PRE_), además de 2 variables con los intervalos de confianza de la media (LMCI_ y UMCI) y del 95% del intervalo central de los valores (LICI y UICI). Así vemos que para el caso 62 con valores de PTH basal de 100 y sin IRC, el valor pronosticado es una PTH de 103.2, con un intervalo del confianza del 95% entre 65.5 y 141.9; el 95% de los pacientes van a presentar un valor de PTH entre -93.97 y 300.40. Como el valor negativo no tendría sentido, este valor se igualaría a cero.

PRE_4	LMCI_1	UMCI_1	LICI_1	UICI_1
103,21640	64,49262	141,94018	-93,97202	300,40482
73,97720	31,43597	116,51844	-123,99628	271,95068
188,51951	153,68033	223,35869	-7,94299	384,98200
168,36725	132,72867	204,00583	-28,23858	364,97308
145,44525	105,21622	185,67428	-52,04429	342,93479
144,35372	103,82591	184,88153	-53,19689	341,90434
132,98823	89,00539	176,97107	-65,30003	331,27649
78,42516	36,82197	120,02835	-119,34887	276,19920
67,91925	23,92406	111,91443	-130,37175	266,21024
155,40540	117,59157	193,21924	-41,60635	352,41715
73,78619	31,20214	116,37024	-124,19650	271,76887
136,57662	93,74824	179,40500	-61,45877	334,61200
87,43023	47,35661	127,50386	-110,02771	284,88818
150,22066	111,22224	189,21909	-47,02188	347,46321
101,07428	62,27284	139,87573	-96,12941	298,27797
163,86471	127,60665	200,12278	-32,85435	360,58378
114,71833	75,82910	153,60756	-82,50265	311,93931
177,50876	142,64359	212,37394	-18,95834	373,97587
128,36238	88,03429	168,69047	-69,14736	325,87212
191,15281	156,17172	226,13390	-5,33490	387,64052

Tras realizar el paso anterior interesa elaborar una tabla con los valores pronosticados según las combinaciones. Para ello, en primer lugar seleccionamos sólo los casos con los nuevos valores que hemos asignados mediante un SELECT IF con selección mediante número de caso ($CASENUM) cuya sintaxis en este caso sería:

```
USE ALL.
COMPUTE filter_$=($CASENUM >= 62).
VARIABLE LABELS filter_$ '$CASENUM >= 62 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

A continuación elaboramos la tabla mediante el procedimiento siguiente:

Analizar → Comparar Medias → Medias. Obtendremos el siguiente cuadro de diálogo:

En el cuadro de "Lista de dependientes" seleccionamos la variable "PRE_" que es el valor que se ha previsto. Posteriormente seleccionamos la variable "PTHprevia" en el cuadro "Lista de independientes"; hacemos Clic en el botón "Siguiente" y añadimos la

siguiente variable, en este caso "IRC_CKD". En el botón opciones seleccionamos tan sólo "Media" en el cuadro de "Estadísticos de casilla". De esta forma se va a calcular la media de la variable "PREV_" en función de las 2 variables predictoras, pero como sólo hay un valor en la casilla de "PREV_", el valor de la media será el mismo valor de la casilla. La sintaxis completa sería la siguiente:

```
MEANS TABLES=PRE_4 BY PTHPrevTx BY IRC_CKD
/CELLS=MEAN.
```

Tras ejecutarla obtendremos una tabla, la cual, una vez editada, tendrá un aspecto igual al siguiente:

Valor de PTH al año previsto

Media

	PTH Previa al trasplante				
IRC	100	200	300	400	500
No	73,8	87,4	101,1	114,7	128,4
Sí	136,6	150,2	163,9	177,5	191,2

La principal limitación que presenta SPSS para el uso de la regresión con finalidad descriptiva o predictiva es que como variable independiente no permite la introducción de variables categóricas con más de 2 categorías. La creación de nuevas variables como hemos comentado en apartados previos para tomar una categoría como referencia, no es válida pues SPSS en cada paso de los procedimientos de selección de variables va a considerar estas subcategorías de manera aislada y no la variable en bloque. Por lo tanto, si tenemos una variable categórica con más de 2 categorías tendremos que recodificar la variable para que tome los valores 0 y 1, siendo el valor 0 el de referencia.

REGRESIÓN LOGÍSTICA BINARIA

La regresión logística binaria se aplica en aquellos estudios donde la variable dependiente es una variable categórica con 2 categorías. Es un modelo estadístico que se aplica en un número mayor de estudios que la regresión lineal, en primer lugar porque en la mayoría de los estudios el objetivo fundamental es ver si se produce o no un evento o una exposición (Sí/No) y porque además una variable cuantitativa siempre se va a poder categorizar en 2 categorías a partir de un punto de corte, mientras que una variable categórica nunca se va a poder convertir en cuantitativa. Por tanto, para una variable dependiente cuantitativa vamos a poder realizar estudios mediante regresión lineal y logística.

En la regresión lineal la ecuación de la recta resultante veíamos que predecía el valor que podía tomar la variable dependiente en función de una combinación de valores de las variables independientes. En el caso de la regresión logística, la ecuación resultante lo que predice son proporciones o probabilidades de ocurrir el evento. Así, si la variable dependiente fuese ser o no hipertenso codificados como 0 y 1 (No/SÍ), la ecuación de regresión logística nos daría la probabilidad de ocurrir 1 (Sí) para una determinada combinación de valores de las

variables independientes. Por tanto el valor resultante de la ecuación variará entre 0 y 1. Cuanto más se acerque a 0 más seguro estaremos que el paciente No es hipertenso y cuanto más se acerque a 1 más seguro que el paciente Sí es hipertenso. Suponiendo que una probabilidad de 0.5 no nos permite definir si el paciente es o no hipertenso, todo valor por encima de 0.5 lo catalogaremos como hipertenso (1) y todo valor por debajo de 0.5 lo catalogaremos como no hipertenso (0). Podemos ver la similitud de la regresión logística con una prueba diagnóstica en la que según el resultado podremos determinar si un paciente presenta o no un evento. Por ello en la regresión logística vamos a poder hablar de conceptos como el de Sensibilidad y Especificidad además de construir curvas ROC como si de una prueba diagnóstica se tratase como veremos más adelante.

Los coeficientes de la ecuación de regresión lineal habíamos visto que se interpretaban como "la cantidad" que variaba la variable dependiente por cada unidad de aumento de la variable independiente. En regresión logística los coeficientes se interpretan de un modo diferente. La ecuación del modelo de regresión es la siguiente:

$$y = \frac{1}{1 + e^{-(B0+B1X)}}$$

Y= variable respuesta, equivalente a la probabilidad o proporción predicha.

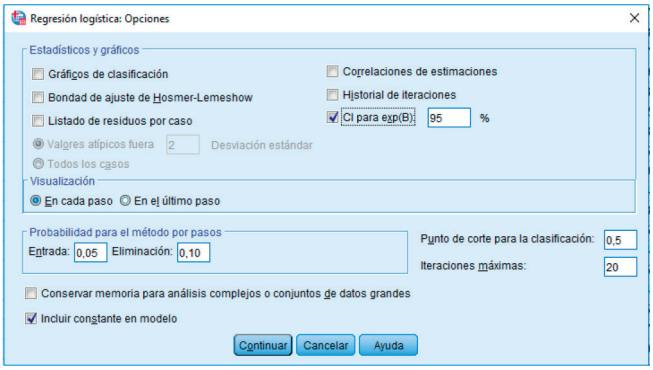
Los coeficientes (B) se van a interpretar como razones de odds u Odds Ratios; es decir, va a ser el valor por el que hay que multiplicar la Odds de la variable dependiente por cada unidad de aumento de la variable independiente.

Vamos a utilizar la base de datos de nuestro ejemplo para ilustrar el modelo de regresión logística. Para ello vamos a estudiar el efecto de la presencia de IRC (variable independiente) en el desarrollo de Acidosis (variable dependiente). El procedimiento sería el siguiente:

Analizar → Regresión → Logística Binaria; se nos abrirá un cuadro de diálogo como el siguiente:

En el cuadro de "Dependientes" seleccionamos la variable dependiente, que en este caso es la presencia de Acidosis (etiquetadas como 0 y 1 como No y Sí). En el cuadro "Covariables" seleccionamos la variable independiente, en este caso presencia de IRC. El campo "Método"

lo modificamos y seleccionamos la opción "Hacia atrás: LR". Hacemos Clic sobre el botón "Opciones" y obtendremos el siguiente nuevo cuadro de diálogo:



Seleccionamos las opciones "CI para exp(B): 95%" para calcular el intervalo de confianza al 95%. La sintaxis completa sería la siguiente:

LOGISTIC REGRESSION VARIABLES EstadAcido
/METHOD=ENTER IRC_CKD
/PRINT=CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

Tras ejecutarla obtendremos los siguientes resultados:

En primer lugar un "Bloque 0" inicial donde se especifican los casos que se han utilizado para la elaboración del modelo de regresión, así como los casos perdidos; además se especifica cómo se han codificado las variables categóricas. En el "Bloque 1" obtenemos las siguientes tablas de resultados:

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	11,523	1	,001
	Bloque	11,523	1	,001
	Modelo	11,523	1	,001

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	69,314[a]	,172	,234

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

Tabla de clasificación[a]

Observado			Pronosticado		
			EstadAcido		Porcentaje correcto
			No	Sí	
Paso 1	EstadAcido	No	23	15	60,5
		Sí	4	19	82,6
	Porcentaje global				68,9

a. El valor de corte es ,500

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	IRC_CKD	1,986	,642	9,551	1	,002	7,283	2,068	25,657
	Constante	-1,749	,542	10,426	1	,001	,174		

a. Variables especificadas en el paso 1: IRC_CKD.

Modelo si el término se ha eliminado

Variable		Logaritmo de la verosimilitud de modelo	Cambio en el logaritmo de la verosimilitud -2	gl	Sig. del cambio
Paso 1	IRC_CKD	-40,419	11,523	1	,001

En la primera tabla "Prueba ómnibus de coeficientes del modelo" obtenemos la significación global del modelo, donde vemos que el Modelo en el paso 1 obtiene un Chi- Cuadrado de 11.523 con una p=0.001. Esta significación se estima mediante el método de Máxima Verosimilitud a través de un Chi-Cuadrado; en la regresión lineal veíamos que la significación del modelo se estimaba mediante la Suma de Cuadrados de los residuales.

Por otra parte, en la regresión lineal obteníamos del valor del coeficiente de determinación R2 que nos informaba del porcentaje de variabilidad total que explicaba el modelo de regresión. Como similitud a este parámetro tenemos en la regresión logística el valor de "R cuadrado de Nagelkerke" que podemos observar en la segunda tabla de "Resumen del Modelo". En este caso el modelo explica un 23.4% de la incertidumbre total de la variable dependiente.

A continuación tenemos otra tabla que se llama "Tabla de Clasificación"; se llama así porque clasifica a los pacientes en función de si el modelo etiqueta a un paciente como acidótico y en realidad lo es, o como no acidótico cuando en realidad no lo es. Vemos que es equivalente a una prueba diagnóstica. Los valores a tener en cuenta serán los de la diagonal porque son los pacientes que están correctamente clasificados. En la última columna de esta tabla aparece el "Porcentaje correcto": el valor de la primera fila 60.5% representa a la Especificidad de la prueba; el valor de la segunda fila 82.6% a la Sensibilidad; y la última fila al % global correctamente clasificado, que en este caso vale 68.9%. Por regla general, se considera que una prueba es adecuada cuando la sensibilidad, especificidad y porcentaje total correctamente clasificado es mayor al 75%.

En este caso, salvo la especificidad, no supera el 75% y por tanto no sería una prueba diagnóstica adecuada.

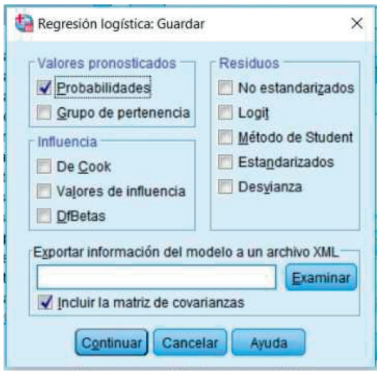
Por último tenemos la tabla de "Variables en la ecuación"; en ella podemos ver el valor de la variable independiente en la columna "Exp(B)" que en este caso es de 7.283; podemos decir que la Odds de acidosis en los pacientes con IRC es 7.28 veces superior al de los pacientes sin IRC, con un IC95% que varía entre 2.07 y 25.66, con una significación p=0.002 según la prueba de Wald. Pero en lugar de utilizar la prueba de Wald para ver la significación del coeficiente, vamos a utilizar la significación a través del logaritmo de la máxima verosimilitud que viene en la última tabla "Modelo si el término se ha eliminado" donde la significación es p=0.001. Si en

el cuadro inicial no hubiésemos seleccionado la opción "Hacia atrás: LR" sino la opción por defecto "Intro" no nos hubiese dado el valor de la significación a través de esta prueba y sólo habríamos tenido la de Wald, por eso es importante hacer la selección correcta al inicio.

Vemos por tanto que este valor es una medida del riesgo de acidosis por cada aumento de una unidad de la variable independiente. Pasar de no tener IRC (valor 0) a tenerla (valor 1) supone 7.28 veces más riesgo de acidosis. Se describiría así para un estudio: OR 7.28 (IC95%: 2.07 a 25.66), p=0.001.

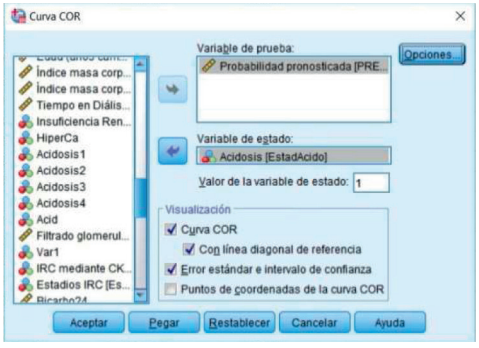
Como hemos podido observar, la regresión logística es similar a una prueba diagnóstica. La forma de valorar este tipo de prueba, además de la sensibilidad y especificidad, es útil la representación gráfica para valorar de manera visual su idoneidad. Esto se realiza mediante las Curvas ROC. Vamos a ver como se realiza:

En el cuadro de diálogo inicial del modelo de regresión, hacemos Clic sobre el botón "Guardar" y seleccionamos la opción "Valores pronosticados → Probabilidades" tal y como vemos en el cuadro de diálogo siguiente:



Esto nos creará una nueva variable en la tabla de datos con el nombre "PRE_" representando los valores pronosticados por el modelo. Para elaborar el gráfico de curva ROC debemos comparar estas probabilidades predichas con las observadas, de la siguiente manera:

Analizar → Curva COR; se nos abrirá el siguiente cuadro de diálogo:



En el cuadro "Variable de prueba" seleccionamos la variable nueva creada con las probabilidades pronosticadas por el modelo "PRE_". La variable de comparación en este caso será la variable Acidosis que son los valores observados, seleccionando el valor 1 en el cuadro "Valor de la variable de estado" porque es el valor con el que hemos etiquetado la presencia de acidosis. En la opción "Visualización" seleccionamos las opciones "Curva COR", "Con línea diagonal de referencia" y "Error estándar e intervalo de confianza". Tras ejecutar la sintaxis obtenemos lo siguiente:

Área bajo la curva

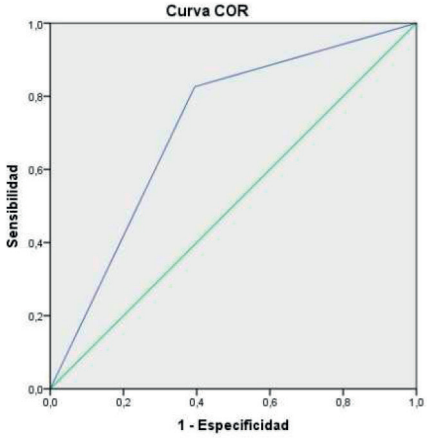
Variables de resultado de prueba: PRE_5

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,716	,067	,005	,584	,848

Las variables de resultado de prueba: PRE_5 tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5



Los segmentos de diagonal se generan mediante empates.

En la primera tabla se especifica el valor del área bajo la curva en la columna "Área", que en este caso vale 0.716, con un IC95% de esta área que varía entre 0.584 y 0.848. Cuanto más se acerque el área a 1 mejor será la prueba. Nota: el IC95% nunca puede ser superior a 1, pero el programa sí es posible que nos dé un valor superior a 1 por la metodología por la que se calcula (método no paramétrico); en este caso a la hora de expresar los resultados en un estudio daríamos el valor superior como 1.

En el gráfico podemos ver representada la sensibilidad y la especificidad de la prueba. Cuanto más se aleje de la diagonal y se acerque al cuadrante superior izquierdo mejor.

En este caso vemos que se aleja de la diagonal, pero no se acerca demasiado al cuadrante superior derecho, por lo que el modelo de regresión no sería muy adecuado (hay que tener en cuenta que tan sólo se ha realizado con una variable independiente).

Para expresarlo en un estudio pondríamos el gráfico y debajo el área de la curva junto con su intervalo de confianza.

Hasta ahora hemos elaborado un modelo de regresión logística con tan sólo una variable independiente categórica binaria. Este tipo de modelo es idéntico a la comparación de proporciones mediante una Tabla 2 x 2 que ya hemos visto. En efecto, si realizamos una Tabla de contingencia con estadísticos Chi-cuadrado obtenemos lo siguiente:

Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Razón de ventajas para IRC_CKD (No / Sí)	7,283	2,068	25,657
Para cohorte EstadAcido = No	1,931	1,282	2,909
Para cohorte EstadAcido = Sí	,265	,102	,687
N de casos válidos	61		

Vemos que los resultados son idénticos al modelo de regresión logística previo.

En el caso de variables independientes cuantitativas la interpretación de los coeficientes es similar al de una variable categórica binaria: el valor por el que hay que multiplicar la Odds de la variable dependiente por cada unidad de aumento de la variable independiente. Si en lugar de utilizar la presencia o no de IRC hubiésemos utilizado el filtrado glomerular como variable independiente, el resultado hubiese sido:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	CKD_EPI	-,055	,018	9,121	1	,003	,946	,913	,981
	Constante	2,510	,990	6,423	1	,011	12,303		

a. Variables especificadas en el paso 1: CKD_EPI.

En este caso el valor del Exp(B) del FG vale 0.946. Como es menor de 1 presenta un "efecto protector": por cada ml/min de aumento del filtrado glomerular disminuye el riesgo de acidosis 0.95 veces, con un IC95% entre 0.91 y 0.98, con un valor p<0.001 según el logaritmo de la verosimilitud. Generalmente los efectos protectores se interpretan peor que los de aumento del riesgo. Para expresar el resultado anterior como factor de riesgo se recodifican las variables al contrario de como estén codificadas; para expresar el resultado como factor de riesgo del ejemplo recodificamos la variable Acidosis como "0 tener acidosis" y "1 no tener acidosis"; el resultado ahora sería:

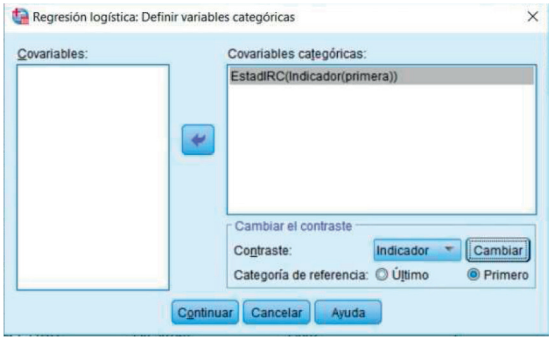
Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	CKD_EPI	,055	,018	9,121	1	,003	1,057	1,020	1,096
	Constante	-2,510	,990	6,423	1	,011	,081		

a. Variables especificadas en el paso 1: CKD_EPI.

Expresándolo como factor de riesgo ahora sería así: por cada ml/min de aumento del FG aumenta el riesgo de No tener acidosis 1.06 veces, con un IC95% entre 1.02 y 1.1 veces.

También podemos utilizar variables independientes categóricas con más de 2 categorías. En el caso de la regresión logística no necesitamos crear nuevas variables para obtener una de referencia como teníamos que hacer en la regresión lineal porque SPSS tiene incorporado en el caso de la regresión logística la creación de nuevas variables de forma automática con el patrón que deseemos (variable de referencia, conjunto de las variables…). Para ello en el cuadro de diálogo inicial haremos Clic sobre el botón "Categóricas". Vamos a verlo utilizando de ejemplo la variable IRC según estadios de la misma. Obtenemos el siguiente cuadro de diálogo al hacer Clic sobre el botón "Categórica":



Seleccionamos la variable categórica en cuestión y la desplazamos hacia el cuadro de la derecha. La opción "Contraste" nos permite seleccionar el tipo de comparación que queremos: Indicator (categoría de referencia, que es la más frecuente), Polinómico (Tendencia Lineal, cuadrática…), etc. Vamos a seleccionar la opción "Indicator" porque es la más utilizada; como "categoría de referencia" vamos a utilizar la "Primera"; seleccionamos "Primero" y hacemos Clic sobre el botón "Cambiar" para que se haga efectivo porque por defecto viene la última categoría como referencia. Tras ejecutar la sintaxis obtenemos:

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetro (1)	(2)
EstadIRC	Sin IRC	27	,000	,000
	Estadio 3	30	1,000	,000
	Estadio 4	4	,000	1,000

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	EstadIRC			9,990	2	,007			
	EstadIRC(1)	1,883	,654	8,293	1	,004	6,571	1,825	23,667
	EstadIRC(2)	2,848	1,275	4,985	1	,026	17,250	1,416	210,120
	Constante	-1,749	,542	10,426	1	,001	,174		

a. Variables especificadas en el paso 1: EstadIRC.

En la primera tabla vemos como se codifican los contrastes, donde la categoría de referencia es la categoría "Sin IRC".

En la segunda tabla obtenemos que: los pacientes con Estadio 3 de IRC (EstadIRC(1)) respecto a los pacientes sin IRC tienen 6.57 veces más riesgo de acidosis. Los pacientes con Estadio 4 de IRC (EstadIRC(2)) respecto de los pacientes sin IRC tienen 17.25 veces más riesgo de acidosis metabólica.

Diagnósticos del modelo de regresión logística

Al igual que sucedía con la regresión lineal, debemos comprobar si el modelo creado es correcto o está influenciado por la presencia de valores alejados de la variable X (predictora), Y (dependiente) o existen valores que influyen de manera notable en la estimación de los coeficientes. En regresión logística lo que se hace es valorar la diferencia entre la probabilidad pronosticada y la observada.

Valores alejados de la variable Y (SRESID)

Valores del residual estudentizado (SRESID) superiores a 2 (en valor absoluto), y sobre todo superiores a 3 nos debe hacer plantear si eliminarlo de la base de datos porque indicaría un valor alejado de la variable respuesta que nos podría condicionar los resultados del modelo elegido.

Valores alejados de la variable X (Valor de Influencia)

Se recomienda valorar la eliminación de los casos cuyo Valor de Influencia sea superior al resultado de la fórmula $2(p+1)/n$, donde "n" es el tamaño de muestra y "p" el número de predictores del modelo. En nuestro ejemplo de Acidosis e IRC el punto de corte estaría en:

n=61

p=1 Por tanto Valor de influencia= $2(1+1)/61= 0.0656$

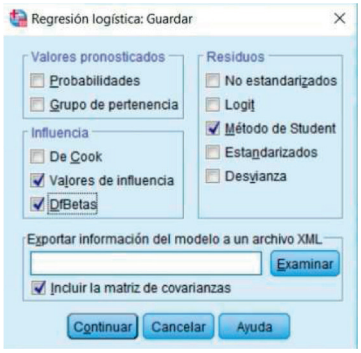
Valores influyentes en la estimación de los coeficientes

En este caso usamos una variante de las Distancias de Cook del modelo de regresión lineal que es el DFBETA. Se recomiendan revisar los valores superiores al punto de corte (en valor absoluto) derivado de la fórmula $2/\sqrt{n}$, donde "n" es el tamaño de muestra. En nuestro ejemplo de Acidosis e IRC el punto de corte estaría en:

n=61

Por tanto DFBETA=$2/\sqrt{n}=0.256$

Para obtener los resultados anteriores, en el cuadro de diálogo de regresión inicial, hacemos Clic sobre el botón "Guardar" y seleccionamos las opciones anteriores, como vemos a continuación:



Para que SPSS nos liste los casos con sus correspondientes valores diagnósticos en una tabla, debemos modificar la sintaxis "a mano" incluyendo la opción "CASEWISE" como vemos a continuación:

LOGISTIC REGRESSION VARIABLES EstadAcido
/METHOD=ENTER IRC_CKD
/CASEWISE= LEVER DFBETA SRESID
/SAVE=LEVER DFBETA SRESID
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

La tabla resultante nos informa sobre los casos que están incorrectamente clasificados, además de sus valores diagnósticos. Parte de la tabla resultante es la siguiente:

Lista por casos

Caso	Estado seleccionado [a]	Observado	Variable temporal			
		EstadAcido	Influencia	SResid	DFB0	DFB1
1	S	N	,037	-,577	-,045	,045
2	S	N**	,029	-1,299	,000	-,069
3	S	S	,029	1,095	,000	,054
4	S	N**	,029	-1,299	,000	-,069
5	S	N**	,029	-1,299	,000	-,069
6	S	S	,029	1,095	,000	,054
7	S	N**	,029	-1,299	,000	-,069
8	S	N	,037	-,577	-,045	,045
9	S	S	,029	1,095	,000	,054
10	S	S	,029	1,095	,000	,054
11	S	N**	,029	-1,299	,000	-,069
12	S	N	,037	-,577	-,045	,045
13	S	N	,037	-,577	-,045	,045
14	S	N	,037	-,577	-,045	,045
15	S	N**	,029	-1,299	,000	-,069
16	S	N	,037	-,577	-,045	,045
17	S	S	,029	1,095	,000	,054
18	S	N	,037	-,577	-,045	,045
19	S	N	,037	-,577	-,045	,045
20	S	S**	,037	1,991	,260	-,260
21	S	S	,029	1,095	,000	,054
22	S	N	,037	-,577	-,045	,045
23	S	S	,029	1,095	,000	,054
24	S	S	,029	1,095	,000	,054
25	S	N**	,029	-1,299	,000	-,069
26	S	S**	,037	1,991	,260	-,260
27	S	N	,037	-,577	-,045	,045
28	S	N	,037	-,577	-,045	,045
29	S	S	,029	1,095	,000	,054

Vemos que no existe ningún Valor de Influencia superior a 0.0656. Ningún valor (en valor absoluto) de SRSID superior a 2. Los casos 20 y 26 tienen valores DFBETA (DFB0) superiores a 0.256 y habría que revisarlos. Por otra parte vemos que hay muchos casos que están mal clasificados (marcados con doble asteriscos **) y nos debería replantear si el modelo es adecuado o no, aunque primero habría que volverlo a estimar tras eliminar o corregir los casos que no cumplen con los criterios diagnósticos anteriores.

El tamaño de muestra es el principal condicionante para la estimación de un modelo con buena capacidad predictiva y que no vulnere los criterios diagnósticos anteriores. Generalmente se exige al menos 10 casos expuestos y no expuestos (si es un estudio de Cohortes) o 10 casos y 10 controles (si es un estudio de Casos-controles) por predictor para tener un tamaño adecuado de muestra.

Al igual que sucedía con la regresión lineal, las aplicaciones de la regresión logística son las mismas, es decir, se va a utilizar para valorar la influencia de una variable independiente sobre la variable dependiente, ajustada por el resto de variables de confusión (método para valorar un efecto) y para valorar cuáles son las diferentes variables que determinan de manera más notable los cambios sobre la variable dependiente (método descriptivo) además de elaborar una ecuación que nos permita predecir la variable dependiente para un conjunto de valores de las variables independientes (método predictivo). Por tanto, los comentarios realizados sobre estos conceptos en la regresión lineal son aplicables a la regresión logística.

Modelo de regresión logística para medir un efecto

Para valorar el efecto de una determinada variable predictora sobre una variable dependiente, debemos evaluar la existencia en primer lugar de variables de confusión y ajustar por ellas. Una variable de confusión como hemos comentado en la regresión lineal, era aquella que se había demostrado en otros estudios que se relacionaba con la variable dependiente en cuestión. En nuestro ejemplo de acidosis e IRC, serían aquellas variables que se hayan demostrado en otros estudios que influyen en la presencia de acidosis independientemente de la presencia o no de IRC. Estas variables deben ser recogidas cuando se realiza un estudio.

Primer criterio

En primer lugar vamos a ver si existe un disbalance entre las variables a estudio entre los grupos con y sin IRC. Para ello realizaremos regresiones logísticas univariantes (en este caso dado que la presencia de IRC es una variable categórica binaria) o regresiones lineales univariantes (si la variable a valorar el efecto es cuantitativa como podría ser el Filtrado Glomerular en ml/min). Vamos a considerar como posibles variables de confusión el Sexo (Sexo), las cifras de PTH al año (PTH4), la dosis de Corticoides (DosisCortico), la Edad (Edad) y el Tiempo en diálisis (TiempHD).

Seleccionaremos como posibles variables de confusión sólo aquellas que sean estadísticamente significativas (o aquellas, que aun no siéndolo, creamos oportuno el ajustar por ellas).

Cuando se trata de una regresión logística, una forma de ver cuáles son o no significativa, en lugar de ir una por una haciendo análisis de regresión logísticas univariantes es observando la tabla "Las variables no están en la ecuación" del Bloque 0; en caso de que la variable a valorar el efecto fuese cuantitativa habría que realizar análisis de regresión lineal una a una, o bien realizar comparación de medias o de proporciones también una a una como vimos en la Regresión Lineal porque no hay forma de obtener una tabla donde aparezcan todas en un solo paso.

Nota: una vez que se sabe hacer regresión lineal y logística, la forma de valorar los disbalances en la variable predictora que se desea valorar el efecto, es mediante estas técnicas porque permiten hacer cualquier tipo de comparaciones, sin necesidad de realizar t-Student, Chi-cuadrado…

En este caso la tabla "Las variables no están en la ecuación" resultante es la siguiente:

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Sexo	4,857	1	,028
		PTH4	6,414	1	,011
		DosisCortico	5,481	1	,019
		Edad	3,878	1	,049
		TiempHD	3,147	1	,076
	Estadísticos globales		18,912	5	,002

Vemos que excepto el Tiempo en diálisis, todas las demás son significativas. El tiempo en diálisis no lo es por muy poco y nos debe hacer pensar si es una variable que interesaría o no introducirla como variable de confusión (a falta de comprobar el segundo criterio) dado su grado de significación.

Segundo criterio

El segundo criterio consiste en valorar si las posibles variables de confusión se relacionan o no con la variable dependiente, independientemente de la variable predictora a estudio. En nuestro caso consiste en ver si las variables previas seleccionadas se relacionan con la presencia de acidosis en el grupo SIN IRC. Teóricamente todas son posibles variables de confusión según el segundo criterio puesto que lo hemos observado en otros trabajos de investigación, pero es posible que haya algunas variables que no esté suficientemente clara su relación y la hayamos recogido por si acaso. Por eso se comprueba este segundo criterio.

Para este segundo criterio primero vamos a seleccionar sólo a los pacientes SIN IRC y realizaremos análisis de regresión logísticas con cada una de las variables. Sólo aquellas con un valor de significación "p" inferior a 0.2 se van a considerar como posibles variables de confusión. Para ello primero seleccionamos a los pacientes SIN IRC mediante el procedimiento:

Datos → Seleccionar casos → Si se satisface la condición → Si… IRC-CKD=0.

De igual manera que en el paso anterior, valoraremos la significación de cada variable a través de la tabla "Las variables no están en la ecuación" del Bloque 0. Vamos a introducir también la variable Tiempo en diálisis porque era casi significativa. El resultado es el siguiente:

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Sexo	1,190	1	,275
		PTH4	,769	1	,380
		DosisCortico	,049	1	,824
		Edad	,765	1	,382
		TiempHD	1,653	1	,199
	Estadísticos globales		3,527	5	,619

En este caso, la variable Tiempo en diálisis es la única con un valor p<0.2. Dado que habíamos decidido introducirla en el segundo paso tras la valoración del primer criterio, sería la única variable por la que habría que ajustar el modelo. Es decir valoraríamos la influencia de la IRC sobre la aparición de Acidosis ajustada por el Tiempo en diálisis.

El modelo final ajustado se valoraría introduciendo ambas variables en el análisis de regresión. Para valorar la significación de manera más precisa de la variable a estudio, se selecciona la opción "Hacia atrás: LR" en el botón "Método" porque de esta forma se valora la significación mediante el logaritmo de la verosimilitud en lugar de mediante la prueba de Wald. El procedimiento lo vemos en el siguiente cuadro de diálogo:

(¡¡¡no olvidar volver a seleccionar todos los casos después del segundo criterio!!!)

LOGISTIC REGRESSION VARIABLES EstadAcido
/METHOD=BSTEP(LR) IRC_CKD TiempHD
/PRINT=CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	IRC_CKD	1,826	,653	7,812	1	,005	6,209	1,726	22,339
	TiempHD	,007	,005	1,729	1	,189	1,007	,997	1,016
	Constante	-2,122	,618	11,775	1	,001	,120		
Paso 2[a]	IRC_CKD	1,986	,642	9,551	1	,002	7,283	2,068	25,657
	Constante	-1,749	,542	10,426	1	,001	,174		

a. Variables especificadas en el paso 1: IRC_CKD, TiempHD.

Como vemos en el paso 1 donde se introducen ambas variables, la presencia de IRC, ajustado al tiempo en diálisis de los pacientes, produce 6.2 veces un aumento del riesgo de acidosis: OR 6.21 (IC 95%: 1.73 a 22.34). La significación la extraemos de la tabla "Modelo si el término se ha eliminado", que en este caso tiene una p=0.003.

Modelo si el término se ha eliminado

Variable		Logaritmo de la verosimilitud de modelo	Cambio en el logaritmo de la verosimilitud -2	gl	Sig. del cambio
Paso 1	IRC_CKD	-38,148	9,103	1	,003
	TiempHD	-34,657	2,121	1	,145
Paso 2	IRC_CKD	-40,419	11,523	1	,001

Por tanto: OR 6.21 (IC95%: 1.73 a 22.34) con una p=0.003.

Modelo de regresión logística con finalidad descriptiva

Al igual que con la regresión lineal, la finalidad en este caso es valorar cuáles son las variables que determinan la aparición de un efecto. Para ello, dentro de un conjunto de variables que hayamos registrado, vamos a seleccionar aquellas que influyan de manera más importante en la aparición del evento observado. Son procedimientos por tanto que se utilizan habitualmente en los estudios retrospectivos de Casos y Controles, donde los pacientes se seleccionan según presenten o no un evento, y se valora de manera retrógrada cuáles son los antecedentes de los pacientes que pudieran haber predispuesto la aparición del evento.

Se debe seguir el mismo principio de parsimonia que vimos en regresión lineal, es decir, dar la máxima información posible con la menor cantidad posible de variables. Para seleccionar las posibles variables, tenemos 2 procedimientos: "Hacia adelante: LR (FSTEP (LR))" o "Hacia atrás: LR (BSTEP (LR))" que son los más apropiados. Los demás procedimientos incorporados en SPSS se aconseja no utilizarlos porque la significación de los parámetros no los da mediante el logaritmo de la máxima verosimilitud que es el más "potente".

Vamos a ver los dos procedimientos suponiendo que los pacientes se hubiesen elegido según la presencia de acidosis o no, y queremos ver qué variables determinan esta presencia de acidosis. Las variables a analizar van a ser: Sexo, Edad, Tiempo en diálisis, Dosis de corticoides, presencia de IRC y niveles de PTH al año.

Método hacia adelante: FSTEP(LR)

Mediante este método primero introduce en el modelo la variable con el menor valor de significación, evaluando de manera externa los no introducidos, introduciendo en el siguiente paso la siguiente variable más significativa. Una vez introducida, valora si alguna de las variables que ya están en el modelo cumplen criterios para salir del mismo (generalmente las que tengan una p>0.1) y por otra parte valora las

variables aún no introducidas, incorporando al modelo aquella que sea más significativa, y así sucesivamente hasta que no haya más variables para introducir (porque tengan una p>0.05) y las que están en el modelo no cumplan el criterio para salir del mismo (porque sus valores "p" son inferiores a 0.1). Veamos el ejemplo en el siguiente cuadro de diálogo:


```
LOGISTIC REGRESSION VARIABLES EstadAcido
/METHOD=FSTEP(LR) Sexo DosisCortico PTH4 Edad TiempHD IRC_CKD
/PRINT=GOODFIT CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Tras ejecutar la sintaxis obtenemos:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	IRC_CKD	1,852	,654	8,020	1	,005	6,375	1,769	22,976
	Constante	-1,658	,546	9,239	1	,002	,190		

a. Variables especificadas en el paso 1: IRC_CKD.

Tras los sucesivos pasos, el modelo sólo incorpora la variable "IRC-CKD". La significación del modelo elegido la obtenemos de la tabla "Prueba ómnibus de coeficientes del modelo" donde vemos que la significación del Modelo tiene una p=0.002.

La significación del término seleccionado (IRC-CKD) la obtenemos de la tabla "Modelo si el término se ha eliminado" que hay justo debajo de la tabla "Variables en la ecuación":

Modelo si el término se ha eliminado

Variable		Logaritmo de la verosimilitud de modelo	Cambio en el logaritmo de la verosimilitud -2	gl	Sig. del cambio
Paso 1	IRC_CKD	-37,048	9,427	1	,002

Por otra parte, además obtenemos la capacidad diagnóstica del modelo seleccionado mediante la tabla "Tabla de clasificación":

Tabla de clasificación[a]

Observado			Pronosticado		
			EstadAcido		Porcentaje correcto
			No	Sí	
Paso 1	EstadAcido	No	21	14	60,0
		Sí	4	17	81,0
	Porcentaje global				67,9

a. El valor de corte es ,500

Tenemos una Especificidad del 60.0% y una Sensibilidad del 81.0%, siendo correctamente clasificados el 67.9% de los casos.

El valor del "R cuadrado de Nagelkerke" es de 0.211, que nos servirá para decidir con qué modelo nos quedamos al final, tras analizar los datos mediante el siguiente método.

Método hacia atrás: BSTEP(LR)

Mediante este método, en primer lugar mete en el modelo todas las variables y posteriormente saca del modelo aquella que sea menos significativa. Una vez sacada esa variable, vuelve a evaluar las variables que quedan en el modelo y saca la que vuelva a ser menos significativa; al mismo tiempo evalúa las variables ya sacadas y vuelve a introducir al modelo aquella que cumpla con el criterio de inclusión ($p<0.05$), y así sucesivamente hasta que las variables del modelo ya no cumplan el criterio para salir del mismo (tener una $p>0.1$) y no haya ninguna variable que esté fuera del modelo que cumpla el criterio para entrar en el mismo (tener una $p<0.05$).

Con las variables del ejemplo anterior se realizaría de la misma manera, pero esta vez seleccionando en "Método" la opción "Hacia atrás: LR". La sintaxis sería:

```
LOGISTIC REGRESSION VARIABLES EstadAcido
/METHOD=BSTEP(LR) Sexo DosisCortico PTH4 Edad TiempHD IRC_CKD
/PRINT=GOODFIT CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5)
```

Tras ejecutarla obtendríamos:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
								Inferior	Superior
Paso 1[a]	Sexo	-,399	,756	,279	1	,597	,671	,153	2,950
	DosisCortico	,000	,000	,112	1	,737	1,000	,999	1,001
	PTH4	,003	,003	1,062	1	,303	1,003	,997	1,009
	Edad	-,044	,029	2,335	1	,127	,957	,905	1,012
	TiempHD	,003	,006	,225	1	,635	1,003	,992	1,014
	IRC_CKD	1,956	,830	5,555	1	,018	7,073	1,390	35,984
	Constante	-,409	1,882	,047	1	,828	,664		
Paso 2[a]	Sexo	-,437	,747	,343	1	,558	,646	,150	2,790
	PTH4	,003	,003	1,249	1	,264	1,003	,997	1,010
	Edad	-,044	,029	2,386	1	,122	,957	,905	1,012
	TiempHD	,004	,005	,650	1	,420	1,004	,995	1,013
	IRC_CKD	2,002	,821	5,949	1	,015	7,402	1,482	36,972
	Constante	-,072	1,588	,002	1	,964	,930		
Paso 3[a]	PTH4	,004	,003	1,283	1	,257	1,004	,997	1,010
	Edad	-,041	,028	2,112	1	,146	,960	,909	1,014
	TiempHD	,003	,005	,483	1	,487	1,003	,994	1,012
	IRC_CKD	1,851	,773	5,734	1	,017	6,368	1,399	28,980
	Constante	-,289	1,551	,035	1	,852	,749		
Paso 4[a]	PTH4	,004	,003	1,404	1	,236	1,004	,998	1,010
	Edad	-,046	,027	2,799	1	,094	,955	,906	1,008
	IRC_CKD	1,978	,758	6,812	1	,009	7,225	1,636	31,898
	Constante	,117	1,454	,006	1	,936	1,124		
Paso 5[a]	Edad	-,051	,028	3,383	1	,066	,951	,901	1,003
	IRC_CKD	2,292	,736	9,694	1	,002	9,893	2,338	41,868
	Constante	,750	1,382	,295	1	,587	2,118		

a. Variables especificadas en el paso 1: Sexo, DosisCortico, PTH4, Edad, TiempHD, IRC_CKD.

Vemos que se han realizado 5 pasos, seleccionando en el último las variables Edad e IRC-CKD. La significación de este último modelo es de una p=0.001 según la prueba ómnibus de coeficientes del modelo. La significación de ambos coeficientes es de 0.053 y <0.001 según la tabla "Modelo si el término se ha eliminado" que hay debajo de la anterior. La capacidad diagnóstica de este modelo seleccionado es el siguiente:

Tabla de clasificación[a]

			Pronosticado		
			EstadAcido		Porcentaje correcto
	Observado		No	Sí	
Paso 1	EstadAcido	No	30	5	85,7
		Sí	9	12	57,1
	Porcentaje global				75,0
Paso 2	EstadAcido	No	30	5	85,7
		Sí	9	12	57,1
	Porcentaje global				75,0
Paso 3	EstadAcido	No	29	6	82,9
		Sí	9	12	57,1
	Porcentaje global				73,2
Paso 4	EstadAcido	No	29	6	82,9
		Sí	10	11	52,4
	Porcentaje global				71,4
Paso 5	EstadAcido	No	29	6	82,9
		Sí	10	11	52,4
	Porcentaje global				71,4

a. El valor de corte es ,500

La Especificidad es del 82.9% y la Sensibilidad es del 52.4%, mientras que el porcentaje de clasificación global es del 71.4%.

Por último el valor del "R cuadrado de Nagelkerke" es de 0.285.

Llegados a este punto, ¿con cuál de los dos modelos nos quedamos? Tendremos que valorar varias cosas. En primer lugar cual es más parsimonioso: el primero sólo tiene un término y por el segundo método obtenemos 2 términos. ¿Tener 2 términos en lugar de 1 mejora mucho la capacidad predictiva del modelo? El primer modelo tiene un R2 de Nagelkerke de 0.211 y el segundo de 0.285 (poca diferencia). Por otra parte ¿cómo es la capacidad diagnóstica de uno y otro? El primero clasifica de manera global de manera correcta a un 67.9% y el último a un 71.4% (ninguno de los dos tiene más del 75%). El segundo tiene mayor Especificad pero a expensas de perder un porcentaje importante de Sensibilidad. Por tanto, nos quedaríamos con el primer modelo que sólo tiene un término.

En última instancia, la selección de un modelo u otro es una decisión subjetiva y será el propio investigador el que decida un modelo u otro según si las variables incluidas en uno u otro le parecen "más convenientes" (por el tipo de variable, por la forma en que se han recogido esas variables, la capacidad para extrapolar los resultados a otras poblaciones, etc.).

Modelo de regresión logística con finalidad predictiva

Con la regresión logística obtenemos una ecuación que nos permite predecir la probabilidad de ocurrencia de un evento (variable dependiente) a partir de un conjunto de combinaciones de valores de las variables predictoras (independientes). Estas probabilidades predichas podemos guardarlas en la tabla de datos como vimos al hablar de la curva ROC. Para predecir la probabilidad a partir de un conjunto de valores que nosotros elijamos, tan sólo tenemos que introducir estos valores en la tabla de datos (sin valor en la variable dependiente) y ejecutar el modelo de regresión que hayamos seleccionado.

Las probabilidades recogidas en la tabla de datos predichas por los modelos de regresión son directamente "Riesgo o Incidencias Acumuladas". Por tanto, si se trata de un estudio de Cohortes o Ensayo Clínico pueden ser utilizadas tal cual. La razón entre dos riesgos por tanto nos daría el Riesgo Relativo. Si tenemos un valor de riesgo que consideramos como referencia, la razón entre el resto de Riesgos y este Riesgo de referencia nos daría el Riesgo Relativo respecto a un valor de referencia. Este valor de referencia generalmente acostumbra a ser aquél que sea más bajo.

En cambio, cuando se trata de un estudio de Casos-Controles, esta probabilidad predicha por el modelo de regresión (Riesgos) debemos reconvertirlos a Odds mediante la fórmula $\frac{PRE\_}{1-PRE\_}$ donde "PRE_" es la probabilidad pronosticada por el modelo anotada en la tabla de datos. La razón entre dos Odds nos daría por tanto la OR.

Las variables que van a formar parte del modelo de regresión con finalidad predictiva se selecciona de igual manera que en el apartado anterior cuando los usamos con finalidad descriptiva. Lo único que en este caso somos menos parsimoniosos y elegiremos aquel modelo que tenga el mejor R2 de Nagelkerke y mejor clasifique a los pacientes según las pruebas diagnósticas. Por tanto, siguiendo con el ejemplo anterior, con finalidad predictiva elegiríamos en modelo con las 2 variables: Edad e IRC-CKD.

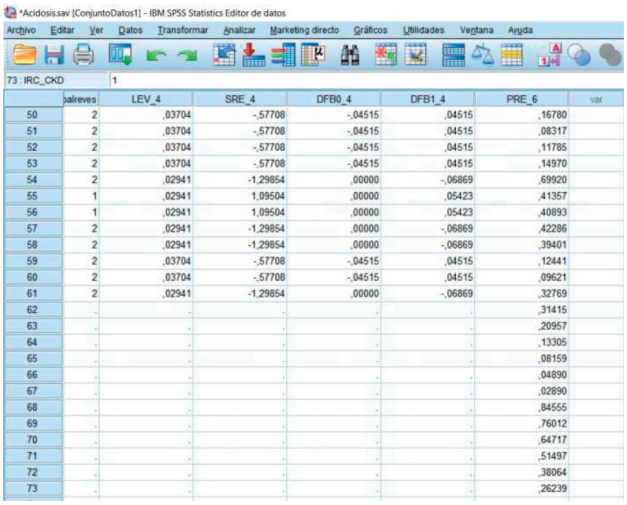
Igual que ocurría con la regresión lineal, es interesante la creación de una tabla donde queden reflejados los RR u OR de cada combinación de variables. Imaginemos que hemos seleccionado el modelo de regresión con las variables Edad e IRC-CKD y queremos crear una tabla con los RR (es un estudio de cohortes) respecto a un valor de referencia. Vamos a ver los riesgos pronosticados para pacientes con edades entre 30 y 80 años con y sin IRC. Deberemos introducir en la base de datos estas combinaciones de valores como vemos en la siguiente imagen:

```
Edad IRC_CKD

30,0     0
40,0     0
50,0     0
60,0     0
70,0     0
80,0     0
30,0     1
40,0     1
50,0     1
60,0     1
70,0     1
80,0     1


Number of cases read:  12    Number of cases listed:  12
```

A continuación ejecutamos de nuevo el modelo de regresión con las dos variables pidiendo que nos guarde los valores de las probabilidades pronosticadas. El resultado sería:



	balreves	LEV_4	SRE_4	DFB0_4	DFB1_4	PRE_6	var
50	2	,03704	-,57708	-,04515	,04515	,16780	
51	2	,03704	-,57708	-,04515	,04515	,08317	
52	2	,03704	-,57708	-,04515	,04515	,11785	
53	2	,03704	-,57708	-,04515	,04515	,14970	
54	2	,02941	-1,29854	,00000	-,06869	,69920	
55	1	,02941	1,09504	,00000	,05423	,41357	
56	1	,02941	1,09504	,00000	,05423	,40893	
57	2	,02941	-1,29854	,00000	-,06869	,42286	
58	2	,02941	-1,29854	,00000	-,06869	,39401	
59	2	,03704	-,57708	-,04515	,04515	,12441	
60	2	,03704	-,57708	-,04515	,04515	,09621	
61	2	,02941	-1,29854	,00000	-,06869	,32769	
62	.	.	.	.	.	,31415	
63	.	.	.	.	.	,20957	
64	.	.	.	.	.	,13305	
65	.	.	.	.	.	,08159	
66	.	.	.	.	.	,04890	
67	.	.	.	.	.	,02890	
68	.	.	.	.	.	,84555	
69	.	.	.	.	.	,76012	
70	.	.	.	.	.	,64717	
71	.	.	.	.	.	,51497	
72	.	.	.	.	.	,38064	
73	.	.	.	.	.	,26239	

Vemos el valor pronosticado de los valores introducidos. Vemos cual es valor del riesgo más bajo, que en este caso es: 0.02890 que corresponde a un paciente de 80 sin IRC. Este va a ser el valor de riesgo de referencia. Para calcular el RR, dividiremos el resto de valores de riesgo entre este riesgo de referencia mediante un COMPUTE.

COMPUTE RR=PRE_ / 0.0289.

EXECUTE.

Al ejecutar esta sintaxis se nos creará la variable RR. Para crear la tabla con los valores de riesgos, primero seleccionamos sólo los pacientes nuevos que hemos introducido al final de la tabla de datos mediante el procedimiento Datos → Seleccionar casos… con la instrucción $CASENUM que identifica a los casos según el número que tienen en la tabla de datos.

USE ALL. COMPUTE filter_$=($CASENUM >= 62). VARIABLE LABELS filter_$

'$CASENUM >= 62 (FILTER)'. VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0).

FILTER BY filter_$.

EXECUTE.

A continuación elaboramos la tabla igual que hicimos en la regresión lineal, a través del procedimiento:

Analizar → Comparar medias → Medias…

MEANS TABLES=RR BY Edad BY IRC_CKD

/CELLS=MEAN.

La tabla resultante (tras editarla) sería:

Informe

Media

	RR					
	Edad					
IRC	30	40	50	60	70	80
No	10,87	7,25	4,60	2,82	1,69	Ref.
Sí	29,26	26,30	22,39	17,82	13,17	9,08

Vemos en la tabla por ejemplo que un paciente de 50 años con IRC tiene 22.39 veces más riesgo de acidosis que el paciente de referencia. El valor de referencia (80 años sin IRC) tiene valor 1, pero en la tabla se ha editado y hemos puesto la notación "Ref." para ilustrar que es el de referencia.

Regresión Logística Multinomial

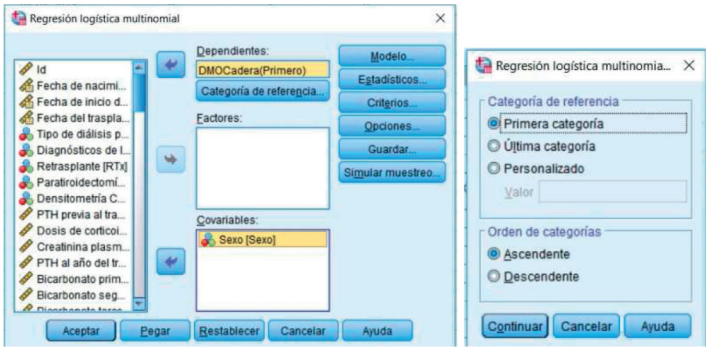
Hasta ahora hemos visto el modelo de regresión logística donde la variable dependiente era binaria (regresión logística binaria), pero en determinados estudios es posible que nuestra variable dependiente no sea binaria sino que sea una variable con más de 2 categorías. Cuando esto sucede tenemos 2 opciones: realizar análisis de regresión logística binaria agrupando las categorías o realizar un análisis de regresión multinomial.

El procedimiento es similar a la regresión logística binaria con la particularidad de que, al igual que sucede cuando tenemos una variable independiente con más de 2 categorías donde elegimos una de referencia, en la regresión multinomial el resultado se interpreta también frente una categoría de referencia. En realidad es como si se hiciesen varias regresiones logísticas binarias de manera simultánea.

Para ilustrar este modelo vamos a tratar de ver en nuestra base de datos la probabilidad de presentar alteración en la densitometría de cadera en los pacientes del estudio. La densitometría de cadera (DMO) la hemos categorizado como 0: normal, 1: osteopenia y 2: osteoporosis. Es decir, vamos a ver cómo influyen las distintas variables sobre la aparición de alteraciones en la DMO de cadera.

En primer lugar vamos a ver el efecto del sexo (variable categórica binaria) sobre la DMO. Para ello vamos a realizar el siguiente paso:

Analizar → Regresión → Logística multinomial; se nos abrirán los siguientes cuadros de diálogo:



En el primer cuadro seleccionamos la variable dependiente, en este caso "DMO de cadera" y la arrastramos al cuadro de "Dependientes". Hacemos Clic sobre el botón "Categoría de referencia" y se nos abre el segundo cuadro de diálogo, donde vamos a seleccionar cuál va a ser nuestra categoría de referencia; en este caso seleccionamos "Primera categoría" dado que vamos a considerar la categoría "0" (normal) como referencia. Tras darle a "Continuar" seleccionamos cual es la variable a estudio, que en este ejemplo es la variable "Sexo". Las variables categóricas se arrastran hacia el cuadro "Covariables". El resto de parámetros los dejamos como vienen predefinidos. La sintaxis es la siguiente:

```
NOMREG DMOCadera (BASE=FIRST ORDER=ASCENDING) WITH Sexo
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20)
LCONVERGE(0) PCONVERGE(0.000001) SINGULAR(0.00000001) /MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE)
ENTRYMETHOD(LR) REMOVALMETHOD(LR) /INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

Tras ejecutarla obtenemos los siguientes resultados:

Información de ajuste de los modelos

Modelo	Criterios de ajuste de modelo: Logaritmo de la verosimilitud -2	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Sólo intersección	16,501			
Final	13,783	2,718	2	,257

Pseudo R cuadrado

Cox y Snell	,044
Nagelkerke	,051
McFadden	,022

Estimaciones de parámetro

DMOCadera[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
Osteopenia	Intersección	,647	,372	3,018	1	,082			
	Sexo	-1,003	,618	2,639	1	,104	,367	,109	1,230
Osteoporosis	Intersección	-,452	,483	,874	1	,350			
	Sexo	-,464	,764	,369	1	,543	,629	,141	2,810

a. La categoría de referencia es: Normal.

En la primera tabla aparece la significación global del modelo de regresión a través del logaritmo de la verosimilitud. En este caso vemos que el modelo no es significativo con una p=0.257. En la segunda tabla aparecen los mismos datos que en la regresión logística binaria con el valor del R2 de Nagelkerke, cuya interpretación es igual.

En la tercera tabla aparecen los valores de los coeficientes de las distintas comparaciones en la columna "Exp(B), junto a su IC95% y la significación estadística, que para la regresión logística multinomial se utiliza la de Wald. Debajo de la tabla aparece una leyenda para recordarnos que "La categoría de referencia es: Normal"¿Cómo se interpretan los resultados? La interpretación puede resultar más dificultosa que en la regresión logística binaria, pero es idéntica. En este caso sería:

- La posibilidad de presentar osteopenia frente a no tener alteraciones óseas (Normal) es 0.367 veces superior para los pacientes con sexo Mujer frente a los hombres, aunque no es significativa (p=0.104).
- La posibilidad de presentar osteoporosis frente a no tener alteraciones óseas es 0.629 veces superior para los pacientes con sexo Mujer frente a los hombres, no siendo en este caso tampoco significativo.

En este ejemplo vemos que el efecto tiende a ser protector (menor de 1). Si queremos expresarlo como factor de riesgo tan sólo hay que recodificar la variable sexo a la inversa.

Si la variable independiente en lugar de tener 2 categorías tuviese más, la interpretación es igual, pero tiene una serie de peculiaridades para su elaboración: en primer lugar, en el primer cuadro de diálogo la variable se debe arrastrar hacia el cuadro "Factores" en lugar de en "Covariables". De esta forma SPSS crea automáticamente las variables ficticias para tomar una de ellas como referencia, pero tomando la categoría más alta de la variable como referencia (al contrario de lo que hacemos habitualmente) y sin poderla modificar. Para tomar la categoría más baja como referencia debemos recodificar la variable. Ejemplo: imaginemos que queremos ver la influencia de la función renal según el estadio de la misma en la aparición de alteraciones en la DMO de cadera. Primero debemos crear una nueva variable de IRC dado que la categoría más baja es la que queremos que sea la referencia (Sin IRC, categoría 1). Para ello le damos a la categoría 1 el valor 6, y al resto los mismos como vemos en la sintaxis:

```
RECODE EstadIRC (1=6) (2=2) (3=3) (4=4) (5=5) INTO IRCregreMul.
EXECUTE.
```

De esta forma la categoría 1 (más baja) pasa a ser la más alta (6). La sintaxis del modelo de regresión sería:

```
NOMREG DMOCadera (BASE=FIRST ORDER=ASCENDING) BY IRCregreMul
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR) REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

Y tras ejecutarla obtenemos:

Pruebas de la razón de verosimilitud

Efecto	Criterios de ajuste de modelo: Logaritmo de la verosimilitud -2 de modelo reducido	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Intersección	15,130[a]	,000	0	.
IRCregreMul	24,763	9,633	4	,047

El estadístico de chi-cuadrado es la diferencia de la log-verosimilitud -2 entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de dicho efecto son 0.

a. Este modelo reducido es equivalente al modelo final porque omitir el efecto no aumenta los grados de libertad.

Pseudo R cuadrado

Cox y Snell	,148
Nagelkerke	,170
McFadden	,078

Estimaciones de parámetro

DMOCadera[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B): Límite inferior	Límite superior
Osteopenia	Intersección	-,442	,427	1,069	1	,301			
	[IRCregreMul=3,00]	1,423	,642	4,916	1	,027	4,148	1,179	14,589
	[IRCregreMul=4,00]	1,540	1,231	1,565	1	,211	4,667	,418	52,121
	[IRCregreMul=6,00]	0[b]	.	.	0	.	.	.	.
Osteoporosis	Intersección	-1,540	,636	5,863	1	,015			
	[IRCregreMul=3,00]	1,828	,835	4,799	1	,028	6,222	1,212	31,937
	[IRCregreMul=4,00]	-16,976	,000	.	1	.	4,240E-8	4,240E-8	4,240E-8
	[IRCregreMul=6,00]	0[b]	.	.	0	.	.	.	.

a. La categoría de referencia es: Normal.

b. Este parámetro está establecido en cero porque es redundante.

En la primera tabla vemos la significación del modelo (p=0.047). En la segunda tabla el valor de Nagelkerke. En la tercera tabla el valor de los coeficientes; en este caso: la probabilidad de osteopenia frente a no tener alteración ósea es 4.15 veces superior en los pacientes con Estadio 3 de IRC frente a los pacientes sin IRC, siendo estadísticamente significativo (p=0.027); mientras que en los pacientes con IRC estadio 4 frente a pacientes sin IRC, la posibilidad de osteopenia es 4.67 veces, pero no es significativo (p=0.211). Por otra parte, la

posibilidad de osteoporosis es 6.22 veces en los pacientes con IRC estadio 3 frente a pacientes sin IRC, con p=0.028; mientras que en los pacientes con Estadio 4 la posibilidad no se ha calculado (vemos un valor del coeficiente prácticamente de 0, sin valor en la significación); la causa de esto último es que no hay ningún paciente en la muestra que presente osteoporosis y Estadio 4 (así como no hay ningún paciente con Estadio 2 no 5 en la muestra global), pero se ha utilizado esta variable como ejemplo.

Si se trata de una variable independiente cuantitativa, la interpretación es igual que en anteriores modelos. En este caso, la variable se debe colocar en el cuadro "Covariables" del cuadro de diálogo inicial (igual que la variable categórica binaria). Veamos el efecto de la dosis de corticoides sobre la presencia de alteración de la DMO de columna. La sintaxis es:

```
NOMREG DMOCadera (BASE=FIRST ORDER=ASCENDING) WITH DosisCortico
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

Tras ejecutarla obtenemos:

Información de ajuste de los modelos

Modelo	Criterios de ajuste de modelo: Logaritmo de la verosimilitud -2	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Sólo intersección	121,322			
Final	116,268	5,053	2	,080

Pseudo R cuadrado

Cox y Snell	,081
Nagelkerke	,092
McFadden	,041

Estimaciones de parámetro

DMOCadera[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B): Límite inferior	Límite superior
Osteopenia	Intersección	-1,132	1,041	1,184	1	,277			
	DosisCortico	,000	,000	1,938	1	,164	1,000	1,000	1,001
Osteoporosis	Intersección	,778	1,673	,216	1	,642			
	DosisCortico	,000	,000	,735	1	,391	1,000	,999	1,001

a. La categoría de referencia es: Normal.

Tenemos un modelo que no llegar a ser significativo con una p=0.080, así como un valor de Nagelkerke muy bajo (0.092). En la última tabla tenemos: por una parte, la posibilidad de osteopenia frente a no tener alteraciones óseas aumenta 1 vez por cada mg de aumento de la

dosis de corticoides, no siendo significativo (p=0.164); por otra parte, la posibilidad de osteoporosis aumenta también una 1 vez por cada mg de aumento de la dosis de corticoides, tampoco siendo significativo con p=0.391.

Hasta ahora sólo hemos introducido una variable predictora, pero es posible incluir variables de ajuste, teniendo la precaución de colocar cada variable en el cuadro correspondiente de "Covariables" y "Factores" del cuadro de diálogo inicial. El resultado será el valor de la variable en cuestión ajustada al resto de variables. Como ejemplo, vamos a ver la influencia de la dosis de corticoides ajustada al sexo de los pacientes. La sintaxis sería:

```
NOMREG DMOCadera (BASE=FIRST ORDER=ASCENDING) BY Sexo WITH
DosisCortico
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001)
SINGULAR(0.00000001)
/MODEL
/STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT=INCLUDE
/PRINT=PARAMETER SUMMARY LRT CPS STEP MFI.
```

El resultado es el siguiente:

Estimaciones de parámetro

DMOCadera[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
Osteopenia	Intersección	-2,079	1,193	3,035	1	,081			
	DosisCortico	,000	,000	2,496	1	,114	1,000	1,000	1,001
	[Sexo=0]	1,156	,649	3,178	1	,075	3,178	,891	11,334
	[Sexo=1]	0[b]	.	.	0	.	.	.	.
Osteoporosis	Intersección	,438	1,735	,064	1	,801			
	DosisCortico	,000	,000	,658	1	,417	1,000	,999	1,001
	[Sexo=0]	,435	,769	,320	1	,571	1,545	,343	6,967
	[Sexo=1]	0[b]	.	.	0	.	.	.	.

a. La categoría de referencia es: Normal.

b. Este parámetro está establecido en cero porque es redundante.

Ajustado al Sexo, la dosis de corticoides no influye en la aparición ni de osteopenia ni de osteoporosis en los pacientes de la muestra.

REGRESIÓN DE COX. ANÁLISIS DE LA SUPERVIVENCIA

En el anterior apartado de regresión logística hemos visto que es posible estimar el riesgo que tiene un sujeto de presentar un evento en función de un patrón de valores que toman las variables independientes; sin embargo, este modelo carece de un factor muy importante a tener en cuenta, que es el factor tiempo. En determinados estudios nos va a interesar no sólo estimar el riesgo que presenta un paciente de presentar un evento sino también evaluar la

"velocidad" a la que aparece ese evento y comparar si esta "velocidad" es diferente según el tipo de paciente.

Mediante el modelo de regresión de riesgos proporcionales de Cox el factor tiempo es tenido en cuenta y estima la tasa instantánea de riesgo de un sujeto, es decir, la tasa de riesgo que tiene un sujeto de presentar el evento en un momento determinado. Inicialmente este modelo de regresión se aplicó para evaluar el riesgo que tenía un sujeto de fallecer y por eso se denominan análisis de supervivencia, pero la variable *exitus* puede ser sustituida por cualquier otra variable que nos interese. La ecuación resultante del modelo de regesión de Cox es la siguiente:

$$h(t;X)= h_0(t) \times e^{Bx}$$

h(t;X)= Tasa instantánea de riesgo

e^{BX}= Función exponencial cuyo exponente es la combinación de las variables independientes.

$h_0(t)$= Función de riesgo de referencia, equivalente a la tasa instantánea de riesgo cuando un sujeto hipotético tenga como exponente el valor 0 en todas las variables independientes.

Cuando comparamos dos sujetos con valores diferentes en las variables independientes mediante regresión de Cox obtenemos la razón de tasas de riesgo, que en análisis de supervivencia se denomina *Hazard Ratio (HR)*. Por ejemplo si hemos catalogado la variable "Diabetes Mellitus" como 0: no presenta DM, y 1: si presenta DM, la razón entre la tasa de riesgo de presentar DM y no presentarla será el riesgo que presenten los pacientes con DM frente a los pacientes sin DM.

$$\frac{h(t;DM=1)}{h(t;DM=0)} = \frac{h0(t)\times e^{B1}}{h0(t)\times e^{B0}} = e^{B1}$$

Para realizar un modelo de regresión de Cox necesitamos recoger tres tipos de variables:

- Variable independiente: es la variable a poner a prueba. Puede ser tanto categórica como cuantitativa.
- Variable dependiente: es la variable respuesta. Será categórica binaria: 0 no presente el evento, 1 Sí presenta el evento. Aunque esta variable la vamos a recoger inicialmente con más de dos categorías.
- Variable tiempo: esta variable será el tiempo transcurrido desde el inicio del estudio y la aparición del evento o la finalización del estudio.

En este tipo de estudio no todos los sujetos se van a incorporar a la vez, sino que se van a ir introduciendo en el estudio a medida que vayan cumpliendo los criterios de inclusión. Por ejemplo, si queremos evaluar la supervivencia de un trasplante renal, los pacientes se irán metiendo en el estudio a medida que se vayan trasplantando. Por tanto, la fecha de inicio del estudio no estará delimitada. Sin embargo, sí debe estar prefijada la fecha fin del estudio. La fecha del fin de seguimiento de un sujeto va a ser diferente según el caso:

- Será la fecha prefijada del fin del estudio si el paciente cumple todo el periodo de seguimiento y no ha presentado el evento.
- La fecha de aparición del evento en caso de presentarlo.
- La fecha en la que se tienen noticias por última vez del paciente porque se haya perdido el seguimiento posteriormente (por traslado a otro hospital, porque deja de venir a la consulta, etc.).
- La fecha en la que presenta otro evento distinto al del objetivo del estudio pero que obligan al paciente a salir del mismo (se denominan tiempos censurados).

Por ejemplo, imaginemos que queremos ver la supervivencia de un trasplante renal (tiempo de funcionamiento del mismo) en función del tipo de inmunosupresión utilizada entre el 01/01/2005 y el 31/12/2016. Recogeremos las siguientes variables:

- Variable dependiente: si el paciente al finalizar su seguimiento lo hace con trasplante funcionante o no.
- Variable independiente: tipo de inmunosupresión utilizada.
- Variable tiempo. Será la diferencia entre la fecha de inclusión en el estudio (a medida que se vayan trasplantando) y:
  - La fecha fin del estudio (31/12/2016) si termina el seguimiento con el trasplante funcionante.
  - La fecha en la que el paciente perdió el trasplante y volvió a diálisis.
  - La fecha en la que el paciente dejó de venir a la consulta por traslado a otro hospital de otra comunidad.
  - La fecha en la que el paciente deja el seguimiento por otro motivo: por ejemplo si el paciente fallece con el injerto funcionante. Se denomina "tiempo censurado".

El tiempo de seguimiento de estos pacientes aunque no hayan presentado el evento a estudio también se tiene en cuenta puesto que durante ese periodo "han estado en riesgo" de presentarlo pero no lo han presentado.

Inicialmente la variable dependiente como se ha comentado anteriormente sólo va a tomar valores 0 (no presenta el evento) y 1 (presenta el evento), pero vamos a darle también valor a las otras causas de fin de seguimiento: por ejemplo 2 (pérdida de seguimiento) y 3 (muerte con trasplante funcionante). Esto nos va a servir para describir posteriormente cuáles son las causas de las pérdidas de seguimiento de los pacientes. Un estudio con más de un 10-15% de pérdidas de seguimiento no es valorable.

El tiempo de seguimiento es importante tenerlo en cuenta. Debe ser lo suficientemente largo como para que dé tiempo a que pueda aparecer el evento. Supongamos que el periodo de estudio del ejemplo anterior sólo fuera de 6 meses; es posible que la mayoría de los sujetos terminen el estudio con el trasplante funcionante puesto que es esperable que la pérdida del

mismo se produzca al menos a partir del primer año del trasplante. De igual modo, es conveniente que los sujetos se introduzcan en el estudio de manera constante; imaginemos que en el ejemplo anterior, el periodo de estudio estipulado fuese de 5 años (tiempo suficiente para ver si se pierde la función del trasplante) pero el 80% de los pacientes registrados se incluyen en los últimos 6 meses del estudio. Es obvio que en ese 80% no habrá dado tiempo a desarrollar el evento.

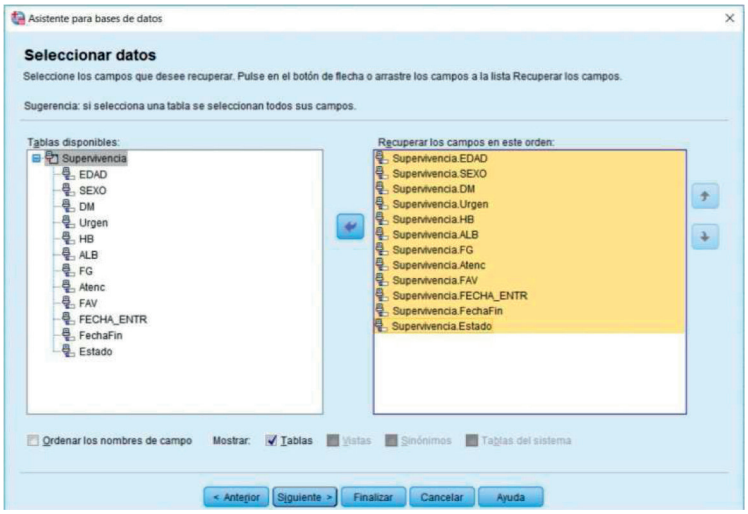
Para ejemplarizar el modelo de regresión de Cox vamos a utilizar la siguiente base de datos, en este caso en formato Excel. El objetivo del estudio es valorar la supervivencia de un paciente que comienza diálisis en función de si comienzan de manera programada o urgente. La base de datos tiene las siguientes variables: edad (en años cumplidos), Sexo (0 mujer, 1 hombre), Diabetes Mellitus (DM: 0 No, 1 Sí), comienzo urgente de diálisis (Urgen: 0 No, 1 Sí), hemoglobina al comenzar, albúmina al comenzar, filtrado glomerular a la entrada, si el paciente es conocido o no previamente en nefrología (Atenc: 0 No, 1 Sí), si tiene fístula arteriovenosa a la entrada (FAV: 0 No, 1 Sí), las fechas de comienzo de diálisis y fin de seguimiento (fecha fin del estudio 1/08/2009) y el estado del paciente al finalizar el seguimiento: 0 vivo, 1 exitus, 2 pérdida de seguimiento, 3 trasplantado.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	EDAD	SEXO	DM	Urgen	HB	ALB	FG	Atenc	FAV	FECHA_ENTR	FechaFin	Estado		
2	47	0	0	1	7,8	2,7	12,67904	0	0	15/01/2002	01/08/2009	0		
3	74	1	0	1	13,2	3,2	4,329077	0	0	31/01/2002	01/08/2009	0		
4	73	0	0	0	12	3,1	6,627866	1	0	05/02/2002	05/02/2006	1		
5	72	0	0	1	9,4	3,1	6,494063	1	0	01/01/2002	11/02/2002	1		
6	77	1	0	1	6,5	3,2	8,846867	0	0	01/01/2002	11/02/2002	1		
7	25	0	0	0	7,2	3,1	4,029524	1	0	26/02/2002	18/07/2004	2		
8	67	0	1	1	11	3,1	6,890746	0	0	15/01/2002	30/01/2002	3		
9	79	1	0	1	7,8	3,1	2,831214	0	0	08/01/2002	04/12/2002	3		
10	69	0	1	1	11,1	3	8,637771	0	0	22/03/2002	01/08/2009	0		
11	88	0	0	0	13	3,5	5,492149	1	0	17/01/2002	06/07/2002	1		
12	46	1	0	1	12,1	1,98	3,587856	0	0	23/01/2002	08/07/2009	2		
13	72	0	0	1	9,7	2,4	4,733497	0	0	03/03/2002	19/03/2002	1		
14	64	0	0	0	14,9	3,1	6,526015	1	1	05/03/2002	01/08/2009	0		
15	70	0	0	0	10,6	2,3	7,010589	1	0	12/03/2002	31/05/2003	3		
16	80	0	0	0	10	2,8	5,09223	1	0	13/03/2002	25/02/2006	1		
17	70	0	1	0	10,2	2,9	6,961242	1	0	21/03/2002	16/03/2004	1		
18	65	0	1	0	12,4	3,6	9,199869	1	0	09/04/2002	03/10/2003	1		
19	73	0	0	1	9,5	3,3	7,042907	0	0	05/04/2003	06/12/2003	1		
20	71	1	0	1	6	2,7	5,004932	1	0	17/04/2002	16/03/2005	1		
21	79	0	1	1	7	3,1	4,344936	0	0	16/04/2002	30/04/2002	1		
22	72	1	1	1	11	2,3	5,586572	0	0	04/05/2002	28/03/2003	3		
23	70	1	0	1	8,3	2,2	5,017969	0	0	10/05/2002	29/05/2002	1		
24	71	0	0	0	11	3,5	6,401614	1	0	24/05/2002	24/07/2002	3		
25	74	0	1	0	9,7	2,8	6,067539	1	0	23/05/2002	01/08/2009	0		
26	50	0	0	1	8	3,5	1,479329	1	0	13/06/2002	09/08/2003	2		
27	66	0	1	1	8	2,5	1,751118	0	0	13/06/2002	01/03/2004	3		
28	29	0	1	1	9,5	2,5	1,245139	1	0	14/06/2002	08/10/2003	3		
29	68	0	1	1	8,4	2,2	8,008123	0	0	21/06/2002	21/11/2007	1		
30	76	1	0	0	10	2,8	5,606623	1	0	26/06/2002	12/03/2007	1		
31	79	1	1	1	10	2,3	17,22816	0	0	15/06/2002	01/06/2006	1		
32	77	1	1	1	10	2,2	6,118961	0	0	31/05/2002	20/01/2003	1		

Hoja1 Hoja2 Hoja3 **Supervivencia**

En primer lugar vamos a importar estos datos desde SPSS. Para ello abrimos SPSS y seleccionamos:

Archivo → Abrir base de datos → Nueva consulta y seleccionamos la opción "Excel Files" marcando la opción "Tablas". Buscamos la hoja de Excel en "Examinar" y tras haberla seleccionado, arrastramos todos los campos a estudio hacia el cuadrado de la derecha como vemos en la siguiente imagen:



Hacemos Clic en "Siguiente" y por último seleccionamos la opción "Pegarlo en el editor de sintaxis para su modificación ulterior". Se nos pegará la siguiente sintaxis:

```
GET DATA
/TYPE=ODBC
/CONNECT='DSN=Excel
Files;DBQ=J:\Curso\Supervivencia.xlsx;DriverId=1046;MaxBufferSize='+
'2048;PageTimeout=5;'
/SQL='SELECT EDAD, SEXO, DM, Urgen, HB, ALB, FG, Atenc, FAV,
FECHA_ENTR, FechaFin, Estado FROM '+ 'Supervivencia'
/ASSUMEDSTRWIDTH=15. CACHE.
EXECUTE.
DATASET NAME ConjuntoDatos1 WINDOW=FRONT.
```

Tras ejecutar la sintaxis, se nos abrirá una tabla de SPSS con los datos. Hay que definir las propiedades de las variables y posteriormente la guardaremos con el nombre "Supervivencia".

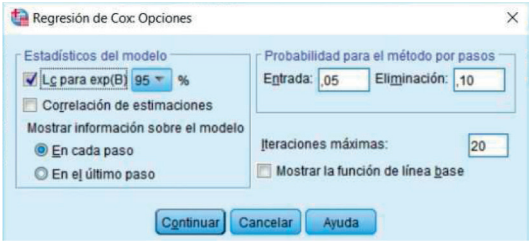
A continuación vamos a calcular el tiempo de seguimiento en meses como vimos al comienzo del curso. La sintaxis sería:

```
COMPUTE Tiempo=CTIME.DAYS(FechaFin - FECHA_ENTR)/30.4375.
EXECUTE.
```

Vamos a valorar la influencia del comienzo urgente de diálisis sobre el riesgo de muerte de los pacientes. Para ello realizamos la regresión de Cox de la siguiente manera:

Analizar → Supervivencia → Regresión de Cox. Se nos abrirá un cuadro de diálogo como el siguiente:

En el cuadro "Hora" vamos a seleccionar la variable creada con el tiempo de seguimiento de los pacientes. En el cuadro "Estado" vamos a seleccionar la variable que nos determina la situación de los sujetos al final de su seguimiento. Seleccionaremos en la opción "Definir evento" la categoría que hayamos predeterminado que corresponde a los eventos, en este caso la categoría la hemos definido como *exitus=1*. En el cuadro "Covariables" vamos a seleccionar la variable que queramos evaluar, en este caso la variable "Urgen". En el botón "Opciones" hacemos Clic y se nos abrirá otro cuadro de diálogo como el siguiente:



Marcamos la opción "Lc para exp(B)95%" en "Estadísticos del modelo" para que nos calcule el IC95% de los parámetros del modelo. La sintaxis completa sería la siguiente:

COXREG Tiempo

/STATUS=Estado(1)

/METHOD=ENTER Urgen

/PRINT=CI(95)

/CRITERIA=PIN(.05) POUT(.10) ITERATE(20).

Y tras ejecutarla obtenemos los siguientes resultados:

Resumen de procesamiento de casos

		N	Porcentaje
Casos disponibles en el análisis	Evento[a]	103	39,6%
	Censurado	155	59,6%
	Total	258	99,2%
Casos eliminados	Casos con valores perdidos	2	0,8%
	Casos con tiempo negativo	0	0,0%
	Casos censurados antes del evento más cercano en un estrato	0	0,0%
	Total	2	0,8%
Total		260	100,0%

a. Variable dependiente: Tiempo

Pruebas ómnibus de coeficientes de modelo[a]

Logaritmo de la verosimilitud -2		1012,643
Global (puntuación)	Chi-cuadrado	8,495
	gl	1
	Sig.	,004
Cambio respecto a paso anterior	Chi-cuadrado	8,327
	gl	1
	Sig.	,004
Cambio respecto a bloque anterior	Chi-cuadrado	8,327
	gl	1
	Sig.	,004

a. Número de bloque de inicio 1. Método = Entrar

Variables en la ecuación

	B	SE	Wald	gl	Sig.	Exp(B)	95,0% CI para Exp(B)	
							Inferior	Superior
Urgen	,571	,199	8,271	1	,004	1,771	1,200	2,614

En la primera tabla se describe el número de casos tenidos en cuenta para el análisis. En este caso hay 2 casos perdidos. Del total de la muestra se ha producido el evento *exitus* en 103 sujetos, que corresponden al 39.6% de la muestra. Es importante fijarse en la información que aparece en "Casos con tiempo negativo": ningún caso puede tener tiempo negativo, si es así, es que alguno de los sujetos tienen las fechas de inicio o final mal registradas en la recogida de datos.

En la segunda tabla se muestra la significación del modelo con la variable "Urgen" en el apartado "Global (puntuación); vemos que el modelo es significativo con una p= 0.004. En la tercera tabla se muestra la influencia de la variable a estudio. Dado que hemos catalogado a los pacientes como 0: no comienzo urgente y 1: comienzo urgente, los resultados nos indican que comenzar diálisis urgente aumenta el riesgo de *exitus* 1.77 veces frente a los pacientes que no comienzan urgente, con un IC95% entre 1.2 y 2.6. Se expresaría así: HR=1.77, IC95%: 1.20 a 2.61.

La significación de esta variable se obtiene a través de la prueba de Wald que en este caso es 0.004.

Si en lugar de una variable independiente categórica binaria, elegimos una variable con más de 2 categorías, elegiremos una categoría como referencia. En este caso, al igual que en regresión logística, SPSS las crea automáticamente haciendo Clic en el botón "Categórica" del cuadro de diálogo inicial. La interpretación de los resultados es igual que en regresión logística.

Lo mismo ocurre para las variables cuantitativas. La interpretación es igual que para la regresión logística. Vamos a poner el ejemplo valorando la influencia de las cifras de hemoglobina (Hb) al comienzo de diálisis. El resultado es el siguiente:

Pruebas ómnibus de coeficientes de modelo[a]

Logaritmo de la verosimilitud -2		1017,875
Global (puntuación)	Chi-cuadrado	3,353
	gl	1
	Sig.	,067
Cambio respecto a paso anterior	Chi-cuadrado	3,441
	gl	1
	Sig.	,064
Cambio respecto a bloque anterior	Chi-cuadrado	3,441
	gl	1
	Sig.	,064

a. Número de bloque de inicio 1. Método = Entrar

Variables en la ecuación

	B	SE	Wald	gl	Sig.	Exp(B)	95,0% CI para Exp(B)	
							Inferior	Superior
HB	-,087	,048	3,348	1	,067	,917	,835	1,006

Vemos que por cada aumento de 1 g/dl de los niveles de hemoglobina el riesgo se multiplica por 0.917 (IC95%: 0.835 a 1.006), p=0.067. En este caso el efecto de la hemoglobina no es significativo (p>0.05 y el IC95% incluye el valor 1), pero la tendencia es a ser inferior a 1 y por tanto a tener un efecto protector; de hecho el límite superior del IC95% es prácticamente 1. Probablemente si aumentásemos el tamaño de la muestra el resultado sería significativo.

De igual modo que sucedía con la regresión lineal y logística, la regresión de Cox también tiene como finalidad el valorar la influencia de una determinada variable ajustada al resto sobre la aparición de un evento, además de las finalidades descriptivas y predictivas.

Modelo de regresión de Cox para medir un efecto

Imaginemos que estamos haciendo un estudio de Cohortes y queremos valorar la influencia del comienzo urgente de diálisis sobre la mortalidad de los pacientes, ajustado a otras variables (variables de confusión) que se han demostrado que también influyen en la supervivencia de estos pacientes. Las variables de confusión a registrar por supuesto deben de haber sido seleccionadas tras haber realizado una exhaustiva revisión bibliográfica sobre el tema que justifique su inclusión en el estudio.

Para evaluar una variable de confusión se deber comprobar los mismos criterios que en el caso de la regresión lineal y logística.

Primer criterio

En este primer criterio vamos a valorar si existe un disbalance de las variables de confusión en los grupos de la variable a estudio. En este caso consistiría en verificar si las variables de confusión están repartidas por igual en los grupos Con y Sin comienzo urgente de diálisis. Para ello realizaremos análisis de regresión logística, lineal, t- Student… entre las variables de confusión y la variable a estudio.

Tras realizar un análisis de regresión logística obtenemos que todas las variables son significativas excepto las variables Edad, Sexo y DM.

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	EDAD	,022	1	,883
		SEXO	1,760	1	,185
		DM	,020	1	,888
		HB	48,389	1	,000
		ALB	17,517	1	,000
		FG	11,569	1	,001
		Atenc	73,092	1	,000
		FAV	49,269	1	,000
	Estadísticos globales		110,231	8	,000

Segundo criterio

En el segundo criterio debemos evaluar si las anteriores variables se relacionan con la mortalidad de los pacientes sólo en el grupo QUE NO COMIENZA URGENTE. Por tanto, en primer lugar habrá que seleccionar sólo a los pacientes con Urgen=0 y realizar un modelo univariante de regresión de Cox con cada una de las variables. Serán variables a tener en cuenta aquellas con una $p<0.2$ para considerarlas como posibles variables de confusión.

Al realizar análisis univariantes con el ejemplo anterior, tan sólo la variable "FAV" presenta un valor "p" a tener en cuenta (p=0.044).

Para considerar una variable como variable de confusión debe cumplir los dos criterios anteriores. En este caso sólo la variable "FAV" los cumple.

Para valorar el efecto de comenzar o no urgente diálisis sobre la mortalidad de los pacientes, ajustado a tener o no fístula arteriovenosa al comienzo (FAV), introducimos ambas variables en el modelo de regresión. El resultado es el siguiente:

NOTA: ¡¡¡No olvidar volver a seleccionar todos los casos!!!

Pruebas ómnibus de coeficientes de modelo[a]

Logaritmo de la verosimilitud -2		1009,078
Global (puntuación)	Chi-cuadrado	9,632
	gl	2
	Sig.	,008
Cambio respecto a paso anterior	Chi-cuadrado	9,769
	gl	2
	Sig.	,008
Cambio respecto a bloque anterior	Chi-cuadrado	9,769
	gl	2
	Sig.	,008

a. Número de bloque de inicio 1. Método = Entrar

Variables en la ecuación

	B	SE	Wald	gl	Sig.	Exp(B)	95,0% CI para Exp(B)	
							Inferior	Superior
Urgen	,404	,222	3,296	1	,069	1,497	,968	2,315
FAV	-,363	,262	1,909	1	,167	,696	,416	1,164

El modelo obtenido es significativo con una p=0.008. El comenzar urgente diálisis sobre no urgente ajustado a tener o no FAV al comienzo, aumenta el riesgo de muerte de los pacientes 1.497 veces, con un IC95% entre 0.968 y 2.315, p=0.069. Es decir, aunque en el análisis univariante el comenzar diálisis de manera urgente aumentaba el riesgo de muerte, tras ajustar por la presencia o no de FAV al comienzo de diálisis pasa a ser no significativo (aunque con tendencia a aumentar el riesgo).

Modelo de regresión de Cox con finalidad descriptiva

Con esta finalidad lo que pretendemos es establecer cuáles son las variables que determinan el desarrollo de un evento, objetivo de los estudios de Casos-Controles. Imaginemos en nuestro caso que se han seleccionado los pacientes según sean exitus o no y queremos ver de manera retrospectiva cuáles son las variables que determinan que los pacientes fallezcan.

Como se ha comentado al principio del capítulo, en este ejemplo se habla de supervivencia, pero el evento a analizar puede ser cualquiera: control de la tensión arterial tras iniciar un tratamiento, cura de un determinado tumor, desarrollo de hemorragia tras iniciar anticoagulación…

Al igual que en los modelos de regresión lineal y logística, SPSS tiene implementado dos formas de seleccionar las variables: mediante un selección por pasos hacia adelante FSTEP o hacia atrás BSTEP.

Método hacia adelante: FSTEP(LR)

Igual que en los modelos de regresión anteriores, en primer lugar SPSS selecciona para el modelo aquella variable que sea más significativa y va introduciendo el resto de variables mientras cumplan el criterio de inclusión que suele ser una p<0.05. Al mismo tiempo en cada paso va analizando las variables que ya están introducidas en pasos previos y saca del modelo aquella que cumpla el criterio de exclusión que suele ser una p>0.1. Realiza estos pasos hasta que ninguna variable cumpla con los criterios de inclusión y exclusión.

Para realizarlo, en el cuadro de diálogo inicial seleccionamos las variables a analizar y en el botón "Método" marcamos la opción "Hacia adelante: LR". La sintaxis sería la siguiente:

```
COXREG Tiempo
/STATUS=Estado(1)
/METHOD=FSTEP(LR) EDAD SEXO DM Urgen HB ALB FG Atenc FAV
/PRINT=CI(95)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

Tras ejecutarla obtenemos los siguientes resultados:

Las variables no están en la ecuación[a]

	Puntuación	gl	Sig.
EDAD	28,748	1	,000
SEXO	,762	1	,383
DM	2,728	1	,099
Urgen	9,150	1	,002
HB	3,046	1	,081
ALB	3,040	1	,081
FG	2,832	1	,092
Atenc	,884	1	,347
FAV	6,191	1	,013

a. Chi cuadrado de residuo = 47,091 con 9 Sig. de gl = ,000

Pruebas ómnibus de coeficientes de modelo[d]

Paso	Logaritmo de la verosimilitud -2	Global (puntuación)			Cambio respecto a paso anterior			Cambio respecto a bloque anterior		
		Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
1[a]	933,032	28,748	1	,000	36,234	1	,000	36,234	1	,000
2[b]	925,201	36,232	2	,000	7,831	1	,005	44,065	2	,000
3[c]	920,795	42,231	3	,000	4,406	1	,036	48,471	3	,000

a. Variables entradas en el número de paso 1: EDAD

b. Variables entradas en el número de paso 2: Urgen

c. Variables entradas en el número de paso 3: FG

d. Número de bloque de inicio 1. Método = Avanzar por pasos (razón de verosimilitud)

Variables en la ecuación

		B	SE	Wald	gl	Sig.	Exp(B)	95,0% CI para Exp(B)	
								Inferior	Superior
Paso 1	EDAD	,054	,010	27,386	1	,000	1,056	1,035	1,078
Paso 2	EDAD	,054	,011	26,513	1	,000	1,056	1,034	1,078
	Urgen	,568	,204	7,747	1	,005	1,764	1,183	2,631
Paso 3	EDAD	,054	,011	26,097	1	,000	1,055	1,034	1,077
	Urgen	,619	,204	9,175	1	,002	1,857	1,244	2,771
	FG	,083	,037	4,986	1	,026	1,086	1,010	1,169

Modelo si se elimina el término

Término eliminado		Pérdida de chi-cuadrado	gl	Sig.
Paso 1	EDAD	36,234	1	,000
Paso 2	EDAD	35,087	1	,000
	Urgen	7,831	1	,005
Paso 3	EDAD	34,653	1	,000
	Urgen	9,264	1	,002
	FG	4,406	1	,036

En la primera tabla vemos la significación de las variables antes de introducirlas en el modelo. Dado que la variable Edad es la más significativa, es la primera en entrar. En la segunda tabla vemos que el modelo final se ha establecido en 3 pasos, con una significación del modelo final $p<0.001$ como podemos ver en el apartado Global (puntuación) de esta tabla. En la tercera tabla vemos las variables seleccionadas en el tercer paso, que en este caso son la

Edad, el comienzo urgente y el Filtrado Glomerular. Es decir, estas son las 3 variables que mejor determinan la supervivencia de los pacientes que comienzan diálisis. La significación de cada variable no la obtenemos a través de la prueba de Wald, sino que la obtenemos mediante el logaritmo de la verosimilitud cuyos resultados aparecen en la tabla "Modelo si se elimina el término".

En este caso por ejemplo vemos que por cada año de edad que aumentan los pacientes, aumenta el riesgo de muerte 1.06 veces, con un IC95% entre 1.03 y 1.08, con una p<0.001. Las variables "Urgen" y "FG" se interpretan de la misma manera.

Método hacia atrás: BSTEP(LR)

En este caso, SPSS primero introduce todas las variables al modelo y posteriormente va sacando variables del mismo mientras cumplan el criterio de exclusión, que suele ser una p>0.1. En cada paso además vuelve a analizar las variables que se han sacado por si alguna de ellas volviera a cumplir el criterio de inclusión (p<0.05) y volvería a introducirla en el modelo.

Para realizarlo el procedimiento es igual al anterior, pero en este caso seleccionando la opción "Hacia atrás: LR" en el botón "Método". La sintaxis sería la siguiente:

```
COXREG Tiempo
/STATUS=Estado(1)
/METHOD=BSTEP(LR) EDAD SEXO DM Urgen HB ALB FG Atenc FAV
/PRINT=CI(95)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

Tras ejecutarla obtenemos los siguientes resultados:

Pruebas ómnibus de coeficientes de modelo[h]

Paso	Logaritmo de la verosimilitud -2	Global (puntuación)			Cambio respecto a paso anterior			Cambio respecto a bloque anterior		
		Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
1[a]	915,917	47,091	9	,000	53,349	9	,000	53,349	9	,000
2[b]	915,997	46,857	8	,000	,080	1	,777	53,269	8	,000
3[c]	916,494	46,454	7	,000	,497	1	,481	52,772	7	,000
4[d]	917,153	45,883	6	,000	,659	1	,417	52,114	6	,000
5[e]	917,923	44,603	5	,000	,770	1	,380	51,344	5	,000
6[f]	919,115	43,260	4	,000	1,192	1	,275	50,151	4	,000
7[g]	920,795	42,231	3	,000	1,680	1	,195	48,471	3	,000

a. Variables entradas en el número de paso 1: EDAD SEXO DM Urgen HB ALB FG Atenc FAV

b. Variable eliminada en el número de paso 2: HB

c. Variable eliminada en el número de paso 3: DM

d. Variable eliminada en el número de paso 4: Atenc

e. Variable eliminada en el número de paso 5: FAV

f. Variable eliminada en el número de paso 6: ALB

g. Variable eliminada en el número de paso 7: SEXO

h. Número de bloque de inicio 1. Método = Pasos sucesivos hacia atrás (razón de verosimilitud)

Pruebas ómnibus de coeficientes de modelo[h]

Paso	Logaritmo de la verosimilitud -2	Global (puntuación)			Cambio respecto a paso anterior			Cambio respecto a bloque anterior		
		Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
1[a]	915,917	47,091	9	,000	53,349	9	,000	53,349	9	,000
2[b]	915,997	46,857	8	,000	,080	1	,777	53,269	8	,000
3[c]	916,494	46,454	7	,000	,497	1	,481	52,772	7	,000
4[d]	917,153	45,883	6	,000	,659	1	,417	52,114	6	,000
5[e]	917,923	44,603	5	,000	,770	1	,380	51,344	5	,000
6[f]	919,115	43,260	4	,000	1,192	1	,275	50,151	4	,000
7[g]	920,795	42,231	3	,000	1,680	1	,195	48,471	3	,000

a. Variables entradas en el número de paso 1: EDAD SEXO DM Urgen HB ALB FG Atenc FAV

b. Variable eliminada en el número de paso 2: HB

c. Variable eliminada en el número de paso 3: DM

d. Variable eliminada en el número de paso 4: Atenc

e. Variable eliminada en el número de paso 5: FAV

f. Variable eliminada en el número de paso 6: ALB

g. Variable eliminada en el número de paso 7: SEXO

h. Número de bloque de inicio 1. Método = Pasos sucesivos hacia atrás (razón de verosimilitud)

En la primera tabla vemos que el modelo final se ha seleccionado tras 7 pasos; la significación del modelo final tiene una $p<0.001$ como podemos ver en el apartado "Global (puntuación).

Abajo, en la tabla de la izquierda, vemos el modelo final seleccionado, en este caso las variables escogidas son "Edad", "Urgen" y "FG". La significación de cada variable la obtenemos de la tabla de la derecha "Modelo si se elimina el término" donde vemos que las significaciones son $p<0.001$, $p=0.002$ y $p=0.036$. La interpretación de cada variable es igual que en el apartado anterior.

En este caso hemos visto que ambos métodos de selección han escogido las mismas variables, pero es posible que no siempre esto ocurra. ¿Con cuál nos quedamos? A ser posible con aquél que sea más parsimonioso, es decir el que contenga menos variable. Otra forma de elegirlo es observando los IC95% de las variables y quedarnos con el que tenga los IC95% más estrechos. Podemos por otra parte seleccionar aquel modelo que "más nos interese" en función de las variables que contengan (viendo qué variables son, si son extrapolables al resto de la población de estudio, si son fáciles de obtener en la práctica clínica diaria,…)

Modelo de regresión de Cox con finalidad predictiva

Al igual que vimos en los anteriores modelos de regresión, una finalidad interesante es la de poder predecir el riesgo de que un sujeto desarrolle el evento de interés en función de los valores que tomen las variables independientes del modelo. Además, es interesante el elaborar una tabla donde aparezcan los Hazard Ratio de un sujeto respecto a uno de referencia.

La selección de las variables con finalidad predictiva es igual que para la descriptiva, siendo menos estrictos a la hora de desechar variables del modelo final.

Mientras que para la regresión lineal y logística, SPSS nos permitía guardar en la

tabla de datos los valores pronosticados, para la regresión de Cox esto no lo hace. Para elaborar la tabla debemos hacerlo a través de los Índices Pronósticos Centrados (IPc). El índice pronóstico del modelo de regresión representa al exponente de la función exponencial del modelo:

Ecuación de modelo de regresión de Cox: $h(t;X)= h_0(t) \times e^{Bx}$

IP=BX

El valor del índice pronóstico del modelo seleccionado en el apartado anterior sería: IP=0.054Edad + 0.619Urgen + 0.083FG

El Índice Pronóstico Centrado representa el valor del índice pronóstico correspondiente a un supuesto sujeto cuyo valor de las variables independientes corresponden a la media de dicha variable. Cuando esto ocurre, el valor del índice pronóstico es igual a 0 representando a un supuesto sujeto promedio. La media de las variables seleccionadas las obtenemos de los resultados de elaborar el modelo de regresión en una tabla al final de los resultados denominada "Medias de Covariables". En el ejemplo anterior es la siguiente:

Medias de covariables

	Media
EDAD	64,394
SEXO	,516
DM	,317
Urgen	,455
HB	10,000
ALB	3,206
FG	6,866
Atenc	,715
FAV	,317

El índice pronóstico centrado para el modelo anterior sería el siguiente:

IPc= 0.054(Edad – 64.394) + 0.619(Urgen – 0.455) + 0.083(FG – 6.866)

Dado que el IPc corresponde al exponente del coeficiente "e" de la ecuación del modelo de regresión, la razón de tasas de incidencias (HR) puede ser expresado a través de los IPC de 2 sujetos A y B:

$$\frac{h(t;A)}{h(t;B)} = \frac{h0(t) \times e^{IPcA}}{h0(t) \times e^{IPcB}} = e^{IPcA-IPcB}$$

Si B representa el sujeto de referencia por ser el sujeto con menor IPc, la fórmula anterior nos dará el HR de cualquier sujeto respecto a un sujeto de referencia.

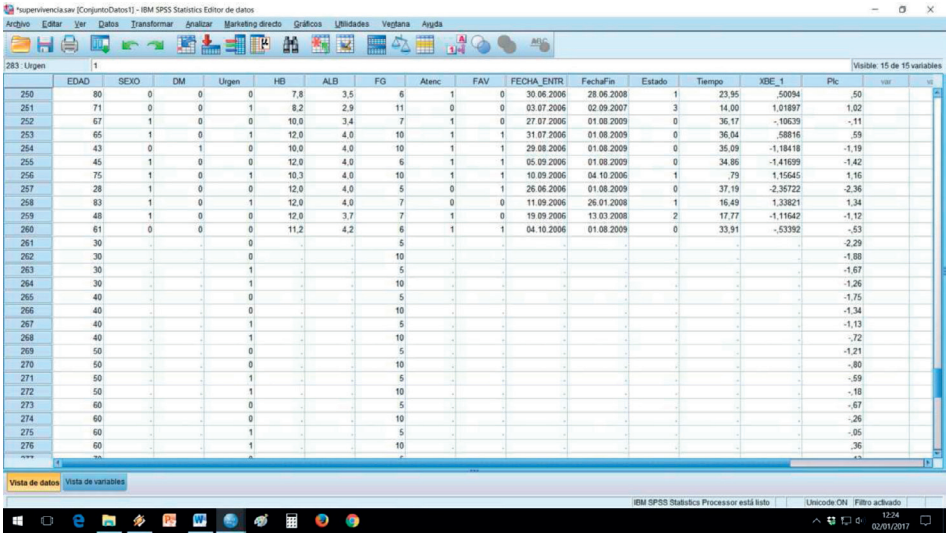
Para elaborar la tabla con los HR, primero debemos introducir "a mano" los patrones de valores en las variables seleccionadas sin introducir valor en la variable tiempo ni en la variable estado. Para el ejemplo anterior vamos a hacerlo para valores de edad 30, 40, 50, 60, 70 y 80, y con FG de 5 y 10 ml/min. La tabla final tendrá el siguiente aspecto:

	EDAD	SEXO	DM	Urgen	HB	ALB	FG	Aten
259	48	1	0	0	12,0	3,7	7	
260	61	0	0	0	11,2	4,2	6	
261	30	.	.	0	.	.	5	
262	30	.	.	0	.	.	10	
263	30	.	.	1	.	.	5	
264	30	.	.	1	.	.	10	
265	40	.	.	0	.	.	5	
266	40	.	.	0	.	.	10	
267	40	.	.	1	.	.	5	
268	40	.	.	1	.	.	10	
269	50	.	.	0	.	.	5	
270	50	.	.	0	.	.	10	
271	50	.	.	1	.	.	5	
272	50	.	.	1	.	.	10	
273	60	.	.	0	.	.	5	
274	60	.	.	0	.	.	10	
275	60	.	.	1	.	.	5	
276	60	.	.	1	.	.	10	
277	70	.	.	0	.	.	5	
278	70	.	.	0	.	.	10	
279	70	.	.	1	.	.	5	
280	70	.	.	1	.	.	10	
281	80	.	.	0	.	.	5	
282	80	.	.	0	.	.	10	
283	80	.	.	1	.	.	5	
284	80	.	.	1	.	.	10	

A continuación creamos la variable IPc con los índices pronósticos de los sujetos. Esto lo hacemos a través de un COMPUTE con la fórmula anterior que hemos elaborado. La sintaxis sería la siguiente:

*COMPUTE IPc=0.054 * (EDAD - 64.394) + 0.619 * (Urgen - 0.455) + 0.083 * (FG - 6.866).*
EXECUTE.

La ejecutamos y obtendremos la variable con los IPc de cada sujeto. Este valor también lo realiza SPSS cuando se ejecuta la regresión de Cox en la opción que aparece en el botón "Guardar" del primer cuadro de diálogo, cuya opción se llama "X * BETA". El problema es que no calcula los IPc de los casos que no tengan valor en la variable "Tiempo" y no nos va calcular el resultado para la combinación de valores que hemos introducido. Podemos pedirle a SPSS que nos guarde los valores de "X * BETA" para comprobar que el COMPUTE que hemos elaborado es correcto comparando el resultado con el resto de sujetos de la tabla. El resultado es el siguiente tras pedir que nos guarde el valor "X * BETA":



Vemos que los resultados son idénticos a los de la variable IPc obtenida con el COMPUTE; vemos también que en los nuevos valores de los patrones introducidos no nos ha creado el valor de "XBE", de ahí que haya que realizarlo a mano como hemos hecho.

A continuación debemos ver cuál es el valor de referencia, que será aquel con el valor del IPc más bajo, pero primero seleccionamos los casos correspondientes al patrón de valores que hemos introducido mediante la opción "$CASENUM > 260". Después pedimos a SPSS que nos diga cuál es el valor más bajo de la variable IPc mediante el procedimiento:

Analizar → Estadísticos Descriptivos → Descriptivos, seleccionando la opción "Mínimo" en el botón "Opciones". El resultado es:

Estadísticos descriptivos

	N	Mínimo
IPc	24	-2,29
N válido (por lista)	24	

Este será el valor de referencia.

A continuación creamos la variable "HR" que será el valor resultante de elevar "e" a la diferencia de cada IPc menos el valor de referencia. Para ello elegimos de las opciones del COMPUTE la opción "EXP" como vemos en el siguiente cuadro de diálogo:

La sintaxis es la siguiente:

COMPUTE HR=EXP(IPc - (- 2.29)).

EXECUTE.

Tras ejecutarla nos habrá creado la variable "HR" para los casos con los patrones de valores que hemos introducido.

A continuación ya sólo nos queda elaborar la tabla igual que hicimos en los anteriores modelos de regresión mediante el procedimiento:

Analizar → Comparar Medias → Medias, pidiendo que sólo nos realice la tabla con el valor de la media (sin el número de caso ni la desviación estándar). No olvidar introducir cada variable dependiente tras hacer Clic en el botón "Siguiente". La sintaxis es:

MEANS TABLES=HR BY EDAD BY Urgen BY FG
/CELLS=MEAN.

Y la tabla resultante, tras editarla es la siguiente:

Informe

Media

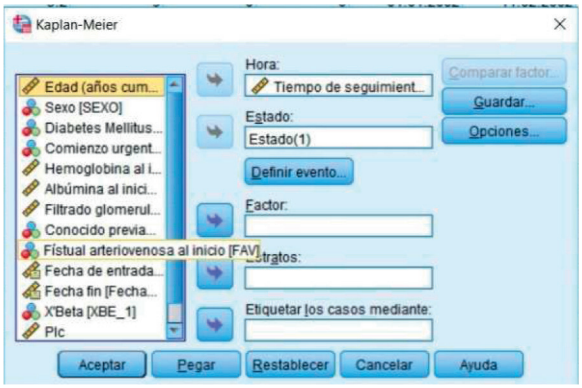
Comienzo urgente	FG (ml/min)	HR					
		EDAD (años)					
		30	40	50	60	70	80
No	5	Ref.	1,71	2,93	5,04	8,64	14,83
	10	1,51	2,59	4,44	7,62	13,08	22,45
Sí	5	1,85	3,18	5,45	9,35	16,05	27,53
	10	2,80	4,81	8,25	14,16	24,30	41,70

Curvas de supervivencia

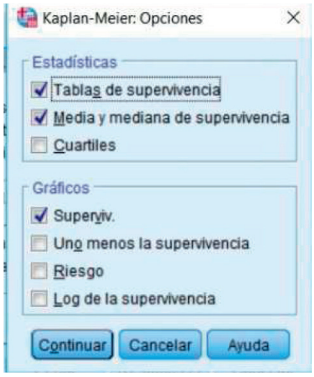
Tras realizar un análisis estadístico mediante un modelo de regresión de Cox, se acostumbra también a realizar una representación gráfica mediante curvas de supervivencia, siendo la curva de Kaplan-Meier la más utilizada. Vamos a ver cómo se realiza, en qué consiste y cómo se interpretan sus resultados.

Vamos en primer lugar a ver la supervivencia de la muestra total. Para realizar la curva de supervivencia tenemos que realizar el siguiente procedimiento:

Analizar → Supervivencia → Kaplan-Meier; se nos abrirá el siguiente cuadro de diálogo:



En el cuadro "Hora" seleccionamos la variable que represente el tiempo de seguimiento de cada sujeto. En el cuadro "Estado" seleccionamos la variable que represente el estado de los sujetos al final del seguimiento. En el botón "Definir evento" seleccionamos el valor de la categoría que represente la aparición del evento, en este caso el valor 1. Hacemos clic en el botón "Opciones" y marcamos las siguientes opciones:



La sintaxis es la siguiente:

```
KM Tiempo
/STATUS=Estado(1)
/PRINT TABLE MEAN
/PLOT SURVIVAL.
```

Tras ejecutarla obtenemos los siguientes resultados:

	Hora	Estado	Proporción acumulada que sobrevive en el tiempo		N de eventos acumulados	N de casos restantes
			Estimación	Error estándar		
1	,427	Exitus	,996	,004	1	258
2	,460	Exitus	,992	,005	2	257
3	,493	Trasplantado	.	.	2	256
4	,526	Exitus	.	.	3	255
5	,526	Exitus	,985	,008	4	254
6	,526	Trasplantado	.	.	4	253
7	,624	Exitus	,981	,009	5	252
8	,657	Exitus	,977	,009	6	251
9	,756	Trasplantado	.	.	6	250
10	,756	Trasplantado	.	.	6	249
11	,789	Exitus	,973	,010	7	248
12	,953	Exitus	,969	,011	8	247
13	1,084	Exitus	,965	,011	9	246
14	1,216	Trasplantado	.	.	9	245
15	1,248	Exitus	.	.	10	244
16	1,248	Exitus	,957	,013	11	243
17	1,347	Exitus	.	.	12	242
18	1,347	Exitus	,949	,014	13	241
19	1,511	Exitus	,945	,014	14	240
20	1,676	Exitus	.	.	15	239
21	1,676	Exitus	,937	,015	16	238
22	2,004	Trasplantado	.	.	16	237
23	2,070	Exitus	,933	,016	17	236
24	2,168	Exitus	,929	,016	18	235
25	2,333	Trasplantado	.	.	18	234
26	2,431	Trasplantado	.	.	18	233
27	2,464	Exitus	,926	,016	19	232
28	2,628	Exitus	,922	,017	20	231
29	2,694	Exitus	.	.	21	230
30	2,694	Exitus	,914	,018	22	229
31	2,891	Exitus	,910	,018	23	228

Una primera tabla en la que aparecen ordenados todos los casos por el orden según han ido ocurriendo los eventos. En esta tabla podemos ver que el primer *exitus* se produjo a los 0.427 meses de seguimiento y el segundo a los 0.460 meses. En la columna "Estimación" aparece la probabilidad acumulada de supervivencia, que representa la probabilidad de que un paciente "sobreviva" más allá de ese periodo. Así podemos ver que el 99.6% de los sujeto sobreviven más allá de los 0.427 meses y que el 96.5% sobreviven más allá del primer mes. Vemos que los pacientes que salen del estudio sin presentar

el evento no tienen asignada ninguna probabilidad de supervivencia puesto que han salido del estudio sin presentar el evento. En la columna "N de eventos acumulados" aparecen los eventos que han ocurrido hasta una fecha concreta; podemos ver que a los 2.628 meses de seguimiento se han producido 20 *exitus* y que el número de sujetos que quedan en riesgo son 231 como podemos ver en la columna "N de casos restantes". En esta última columna se han quitado los pacientes que han presentado el evento y los que han salido del estudio por cualquier motivo.

También aparece la siguiente tabla:

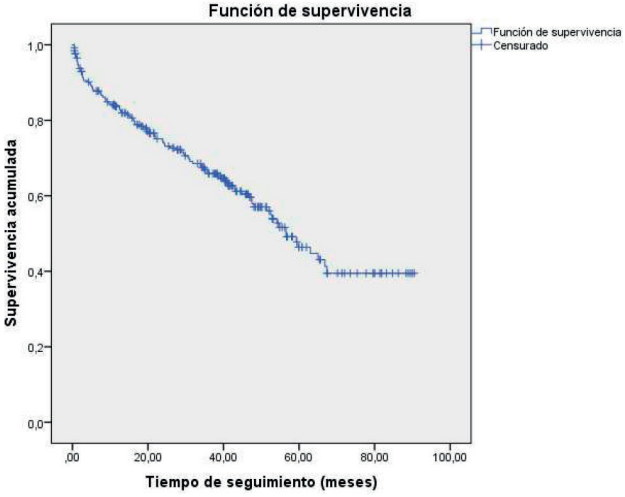
Medias y medianas para el tiempo de supervivencia

Media[a]				Mediana			
		Intervalo de confianza de 95 %				Intervalo de confianza de 95 %	
Estimación	Error estándar	Límite inferior	Límite superior	Estimación	Error estándar	Límite inferior	Límite superior
54,941	2,515	50,011	59,871	56,575	4,305	48,137	65,013

a. La estimación está limitada al tiempo de supervivencia más largo, si está censurado.

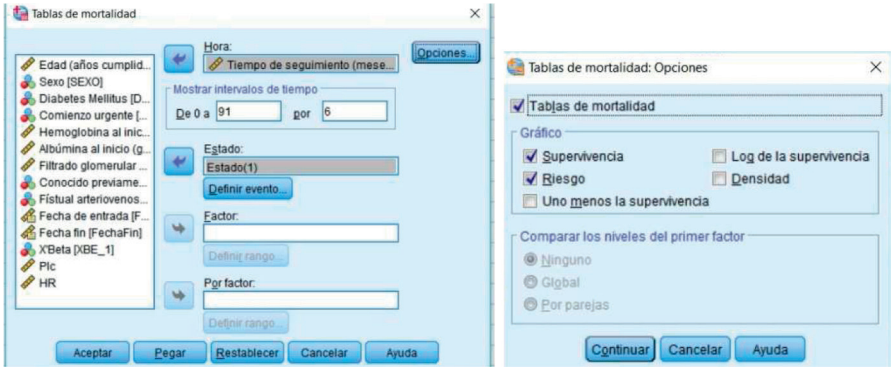
En ella se representa el tiempo medio y mediano de supervivencia de la muestra. Vemos que la media de supervivencia ha sido de 54.94 meses y la mediana de 56.58 meses, con sus respectivos IC95%. En estudios de supervivencia, dado que hay sujetos que salen del estudio sin haber presentado el evento (se pierden por el camino, salen del estudio por otro motivo, etc.) y por tanto sin saber en qué momento hubieran podido presentarlo para disponer de todo el tiempo completo, no es posible calcular con exactitud la media de supervivencia y por ello siempre se acostumbra a poner la mediana. Por tanto en este estudio, la mediana del tiempo de supervivencia es de 56.58 meses (IC95% de 48.14 y 65.01 meses). Esta mediana representa el tiempo en el que el 50% de los pacientes aún están "vivos". Es decir, para este ejemplo, a los 56.57 meses de seguimiento, el 50% de los sujetos aún están vivos.

Por último se representa el gráfico de la supervivencia de los sujetos.



Existe otra forma de evaluar la supervivencia que aporta más información que el método de Kaplan-Meier, que es a través de Tablas de supervivencia o Método Actuarial. Mediante este método, además de los resultados anteriores, nos permite dividir el tiempo de seguimiento en intervalos y ver en cuál o cuáles de ellos se produce la mayor cantidad de eventos. Para realizarlos debemos seguir el siguiente proceso:

Analizar → Supervivencia → Tablas de mortalidad. Se nos abrirá el siguiente cuadro de diálogo:



En el cuadro "Hora" seleccionamos la variable que represente el tiempo de seguimiento. A continuación definimos el tiempo total de seguimiento de la muestra y los intervalos en los que los queremos dividir en la opción "Mostrar intervalos de tiempo". En este caso hemos puesto que el tiempo de seguimiento es de 0 a 91 meses, divididos en intervalos de 6 meses. En el cuadro "Estado" seleccionamos la variable que represente el estado de los sujetos al final de su seguimiento. Hacemos Clic en el botón "Opciones" y se nos abrirá el segundo cuadro de diálogo donde marcaremos las opciones "Tablas de mortalidad", "Gráfico de Supervivencia" y "Gráfico de Riesgo". La sintaxis completa es la siguiente:

```
SURVIVAL TABLE=Tiempo
/INTERVAL=THRU 91 BY 6
/STATUS=Estado(1)
/PRINT=TABLE
/PLOTS (SURVIVAL HAZARD)=Tiempo.
```

Y tras ejecutarla obtenemos los siguientes resultados:

Advertencias

El límite superior de intervalo 91 ha cambiado a 96.

En primer lugar una primera advertencia donde nos dice que SPSS ha cambiado el límite superior de seguimiento a 96 para que "cuadren" los intervalos de 6 meses.

A continuación otra tabla con la siguiente información:

Tabla de mortalidad[a]

Hora de inicio del intervalo	Número que entra en el intervalo	Número de retirada durante el intervalo	Número expuesto a riesgo	Número de eventos terminales	Proporción que termina	Proporción que sobrevive	Proporción acumulada que sobrevive al final del intervalo	Error estándar de la proporción acumulada que perdura al final del intervalo	Densidad de probabilidad	Error estándar de la densidad de probabilidad	Índice de riesgo	Error estándar del índice de riesgo
0	259	9	254,500	31	,12	,88	,88	,02	,020	,003	,02	,00
6	219	14	212,000	10	,05	,95	,84	,02	,007	,002	,01	,00
12	195	9	190,500	12	,06	,94	,78	,03	,009	,002	,01	,00
18	174	13	167,500	9	,05	,95	,74	,03	,007	,002	,01	,00
24	152	9	147,500	7	,05	,95	,71	,03	,006	,002	,01	,00
30	136	7	132,500	9	,07	,93	,66	,03	,008	,003	,01	,00
36	120	27	106,500	5	,05	,95	,63	,03	,005	,002	,01	,00
42	88	14	81,000	6	,07	,93	,58	,04	,008	,003	,01	,01
48	68	16	60,000	4	,07	,93	,54	,04	,006	,003	,01	,01
54	48	11	42,500	6	,14	,86	,47	,04	,013	,005	,03	,01
60	31	5	28,500	2	,07	,93	,43	,05	,005	,004	,01	,01
66	24	4	22,000	2	,09	,91	,39	,05	,007	,004	,02	,01
72	18	4	16,000	0	,00	1,00	,39	,05	,000	,000	,00	,00
78	14	7	10,500	0	,00	1,00	,39	,05	,000	,000	,00	,00
84	7	6	4,000	0	,00	1,00	,39	,05	,000	,000	,00	,00
90	1	1	,500	0	,00	1,00	,39	,05	,000	,000	,00	,00

a. La mediana del tiempo de supervivencia es 57,3320

Vemos debajo de la tabla una leyenda que pone que la mediana del tiempo de supervivencia ha sido de 57.33 meses. Mayor que el que habíamos obtenido con Kaplan- Meier debido a que ha ampliado el seguimiento a 96 meses en lugar de a 91; por tanto la mediana del tiempo de supervivencia la deberemos obtener del Kaplan-Meier.

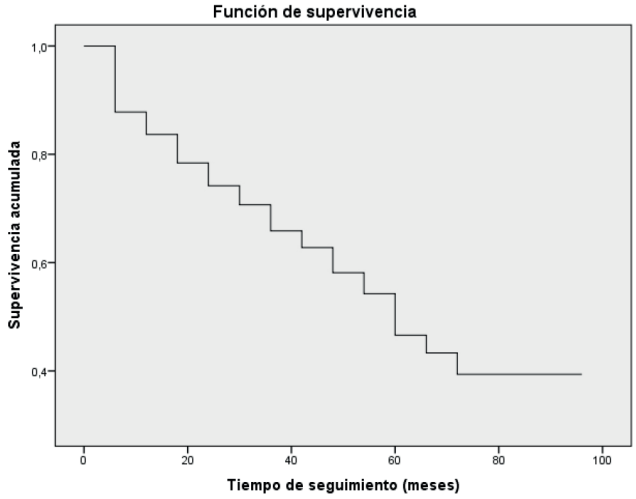
Vemos que al comienzo de los primeros 6 meses de seguimiento (primer intervalo que aparece con el valor 0 en la primer columna), entran en riesgo 259 sujetos. En ese intervalo salen 9 (segunda columna) por distintos motivos, siendo por tanto el número de sujetos expuestos a riesgo (tercera columna) de 254.5 sujetos. Durante este periodo se producen 3 exitus (columna "Número de eventos terminales") representando el 12% del total a riesgo (columna "Proporción que termina"). El 88% restante (columna "Proporción que sobrevive") permanecen vivos y entran en el siguiente intervalo.

En el intervalo de 60 a 66 meses de seguimiento (columna 1 definido por "Hora de inicio del intervalo=60), vemos que sólo quedan 31 sujetos que entran en dicho intervalo. De estos 31, 5 se retiran por distintos motivos, siendo por tanto el número de sujetos a riesgo durante este intervalo de 28.5 sujetos. En este intervalo se producen 2 eventos, que representan el 7% del total de sujetos en riesgo durante este intervalo, mientras que el 93% de ellos sobreviven a dicho intervalo.

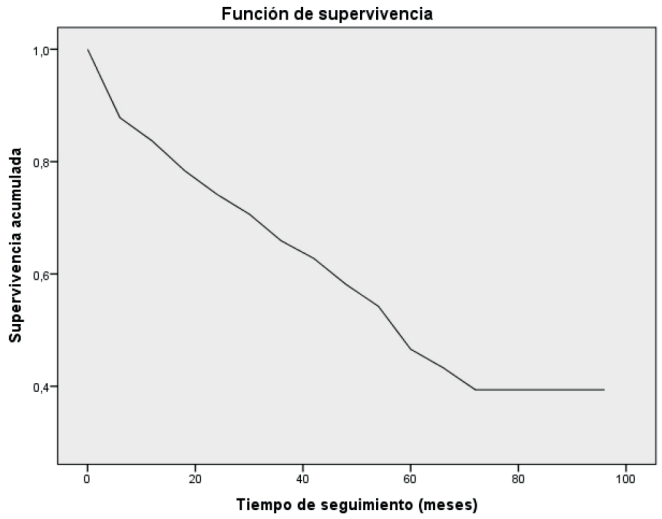
En la octava columna "Proporción acumulada que sobrevive al final del intervalo" aparece el porcentaje del total de sujetos de la muestra que permanecen vivos al final de cada intervalo. Así, al final del intervalo de 60 a 66 meses, el 43% de los sujetos aún están vivos.

En la columna "Índice de riesgo" se representa la tasa relativa media de incidencia de eventos, que representa la proporción de eventos no del total de la muestra, sino de los que llegan vivos a ese intervalo. Nos ayuda a ver en qué intervalo se producen más eventos. Vemos que la mayoría de los exitus se producen en el primer intervalo, pero luego existe un repunte en el intervalo de 56 a 60 meses.

Por último se representan el gráfico de supervivencia:

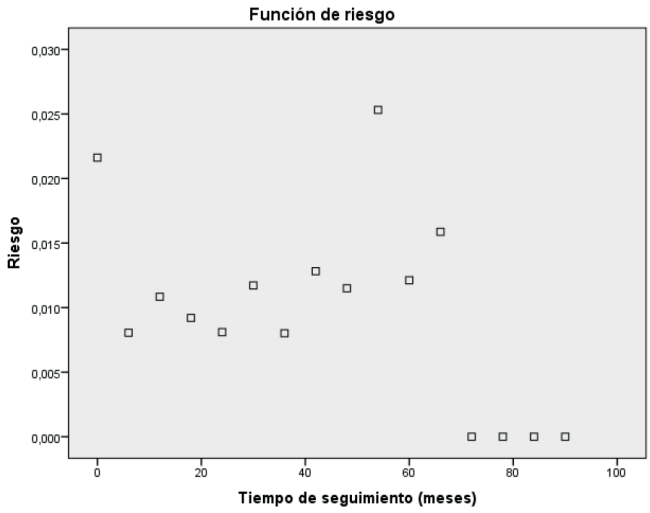


Vemos que es igual que el obtenido mediante Kaplan-Meier. Podemos editarlo pidiéndole a SPSS que nos dibuje una recta en lugar de escalones, y se verá así:

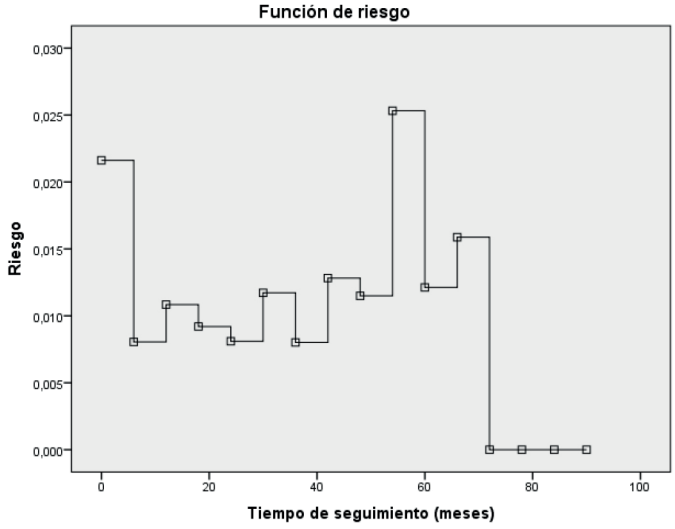


Más parecido al de Kaplan-Meier.

Y por último el gráfico de Riesgo:



El cual vamos a editar, haciendo doble Clic en el gráfico y posteriormente seleccionando los puntos haciendo doble Clic sobre uno de ellos; después hacemos Clic en el botón de la barra de herramientas "Añadir una línea de interpolación" tipo "Paso", "Paso Izquierdo". El resultado es:



Representa la función del índice de riesgo. Tal como comentamos anteriormente, el mayor porcentaje de *exitus* se producen en el primer intervalo, con un repunte posterior en el intervalo 54-60 meses, ahora visto de manera gráfica.

A través del Método de Kaplan-Meier también es posible dibujar sobre un mismo gráfico la supervivencia de la muestra en función de una serie de variables. Además es posible comprobar si las curvas obtenidas son estadísticamente significativas. Esta significación se evalúa a través del logaritmo del rango (LogRank) de Mantel-Haenszel. Este estadístico compara el número de eventos en cada grupo con el número de eventos esperados si la supervivencia de ambos grupos fuera la misma.

Imaginemos que queremos ver la supervivencia de los pacientes del ejemplo anterior en función de si comienzan o no diálisis a través de una FAV. Para ello, en el primer cuadro de diálogo de Supervivencia con Kaplan-Meier, seleccionamos en el cuadro "Factor" la variable por la que se quiere comparar, en este caso "FAV". Posteriormente hacemos Clic en el botón "Comparar factor" y seleccionamos la opción "Log Rango". La sintaxis completa es la siguiente:

```
KM Tiempo BY FAV
/STATUS=Estado(1)
/PRINT TABLE MEAN
/PLOT SURVIVAL
/TEST LOGRANK
/COMPARE OVERALL POOLED.
```

Tras ejecutarla obtenemos los siguientes resultados:

Una primera tabla resumen del número de pacientes en cada grupo y el número y porcentaje de eventos en cada uno de ellos:

Resumen de procesamiento de casos

FAV	N total	N de eventos	Censurado	
			N	Porcentaje
No	174	79	95	54,6%
Sí	83	24	59	71,1%
Global	257	103	154	59,9%

A continuación una tabla como la inicial, pero esta vez dividida en 2 bloques de comienzo a través de FAV No y Sí.

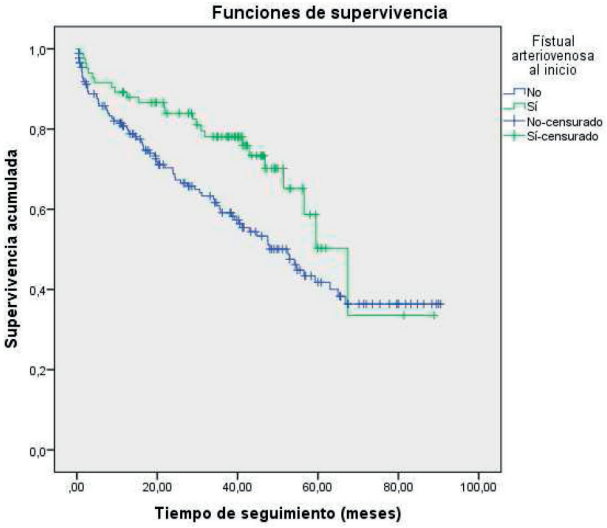
A continuación el tiempo de supervivencia medio y mediano de cada subgrupo y del total de la muestra:

Medias y medianas para el tiempo de supervivencia

FAV	Media[a]				Mediana			
			Intervalo de confianza de 95 %				Intervalo de confianza de 95 %	
	Estimación	Error estándar	Límite inferior	Límite superior	Estimación	Error estándar	Límite inferior	Límite superior
No	50,826	3,028	44,891	56,761	52,468	5,806	41,089	63,847
Sí	58,984	4,911	49,358	68,610	67,417	6,698	54,289	80,545
Global	54,643	2,527	49,691	59,595	56,575	4,255	48,234	64,916

a. La estimación está limitada al tiempo de supervivencia más largo, si está censurado.

Al final de los resultados, se dibuja el gráfico con las supervivencias de cada grupo de sujetos a lo largo del seguimiento:



La significación estadística de la diferencia de ambas curvas de supervivencia aparece en la tabla "Comparaciones Globales":

Comparaciones globales

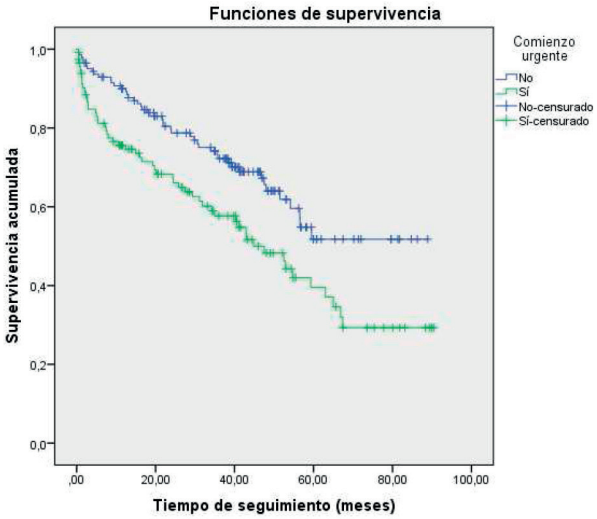
	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	5,948	1	,015

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de FAV.

Podemos ver que la supervivencia de ambos grupos son diferentes, siendo esta diferencia estadísticamente significativa, con una p=0.015. Para representarlo para un estudio se pondría dentro del gráfico el texto: LogRank=0.015.

Si observamos el gráfico, vemos que esta diferencia se produce en los primeros meses de seguimiento, mientras que al final del seguimiento ambas curvas se solapan. Es decir, el comenzar a través de una FAV influye en la supervivencia de los sujetos al principio mientras que conforme avanza el tiempo esta diferencia se pierde.

Veamos ahora el siguiente gráfico donde comparamos la supervivencia en función del comienzo urgente o no de diálisis:



Comparaciones globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	8,499	1	,004

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de Urgen.

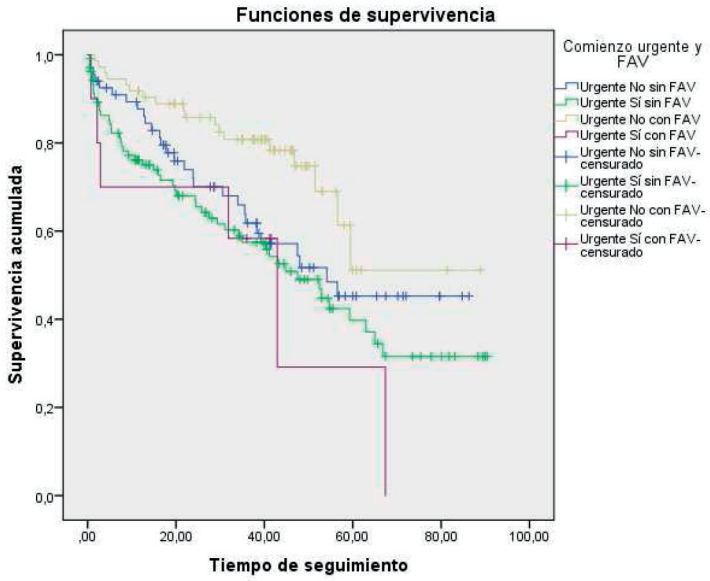
Vemos que ambas curvas también son estadísticamente significativas. Pero tienen una característica especial: esta diferencia se produce fundamentalmente al principio, mientras que después ambas curvas discurren paralelas. Es decir, la influencia del comienzo o no urgente de diálisis se produce sólo al principio del seguimiento y posteriormente se pierde dicha influencia.

Nota importante: Debemos tener presente que las diferencias observadas entre las supervivencias de 2 grupos mediante Kaplan-Meier son análisis brutos que pueden estar artefactados por la presencia de variables de confusión. Únicamente se puede depurar esta confusión mediante análisis de regresión de Cox.

Podemos estratificar las curvas anteriores en función de otra variable arrastrándola hacia el recuadro "Estrato" del primer cuadro de diálogo. Por ejemplo, ver la supervivencia de los pacientes en función del comienzo urgente de diálisis y de si los pacientes tienen o no FAV. Ahora nos hará 2 gráficos de supervivencia, uno para cada estrato. Pero lo ideal es que aparezcan en un mismo gráfico todas las curvas, pero para ello primero tenemos que crear una nueva variable con las posibles opciones a través por ejemplo de un procedimiento IF como el siguiente:

```
IF (Urgen=0 AND FAV=0) Comb=0.
IF (Urgen=1 AND FAV=0) Comb=1.
IF (Urgen=0 AND FAV=1) Comb=2.
IF (Urgen=1 AND FAV=1) Comb=3.
EXECUTE.
```

Tras definir las variables obtenemos el siguiente gráfico de Kaplan-Meier con las 4 curvas de supervivencia:



Posteriormente podemos cambiar el color y las marcas de censura a cada una de las curvas por separado.

ANÁLISIS ESTADÍSTICO PARA PRUEBAS DIAGNÓSTICAS

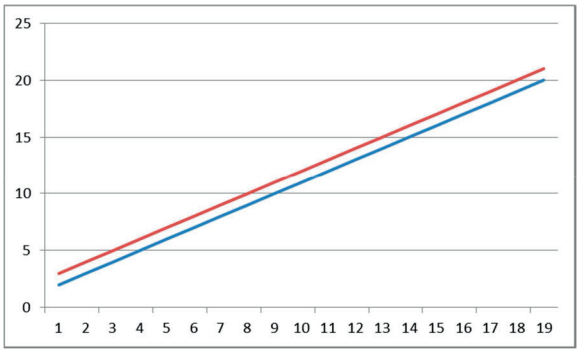
Los estudios basados en pruebas diagnósticas pueden tener distintas finalidades. Una puede ser comparar los resultados obtenidos con una prueba diagnóstica y compararlos con otra que se considera de referencia o Gold estándar por haberse demostrado ser la más fiable hasta la fecha del estudio. En este caso el objetivo del estudio va a ser el validar los resultados obtenidos con la nueva prueba diagnóstica. Por ejemplo, comparar la estimación del filtrado glomerular con la fórmula MDRD y el verdadero filtrado glomerular mediante el aclaramiento de inulina (considerado el método de referencia); o comparar por ejemplo la medición de un tumor mediante una ecografía con la verdadera medición tras la exéresis quirúrgica. Estos serían dos ejemplos donde se comparan dos pruebas a través de variables cuantitativas (ml/min y cm respectivamente), pero otros estudios se compararán a través de variables categóricas. Por ejemplo, la positividad o negatividad de un hemocultivo a través de dos métodos de cultivo distintos.

En otros estudios el objetivo no va a ser el comparar dos pruebas diagnósticas sino el comparar los resultados obtenidos por dos observadores distintos. Por ejemplo, valorar el grado de fibrosis en una muestra de tejido renal por dos patólogos distintos.

Sea cual sea el objetivo del estudio, los análisis estadísticos son los mismos. Vamos a tener análisis estadísticos para pruebas que usan variables cuantitativas y otros para variables categóricas.

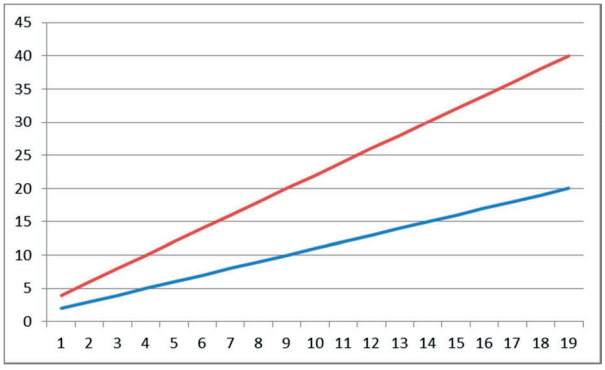
ANÁLISIS ESTADÍSTICOS PARA PRUEBAS DIAGNÓSTICAS CON VARIABLES CUANTITATIVAS

Existen infinidad de artículos publicados donde la comparación de dos pruebas diagnósticas con variables cuantitativas la realizan a través del coeficiente de correlación de Pearson. Sin embargo esto no es correcto puesto que el coeficiente de correlación de Pearson analiza la asociación lineal entre dos métodos pero no nos da información sobre la presencia de diferencias constantes o proporcionales entre los métodos. Como podemos ver en el siguiente gráfico, el coeficiente de correlación de Pearson entre las dos pruebas diagnósticas representadas es 1 (asociación lineal perfecta), sin embargo vemos que existe una diferencia constante entre los dos métodos.



Por ello la forma correcta de evaluarlo es a través del Coeficiente de Correlación

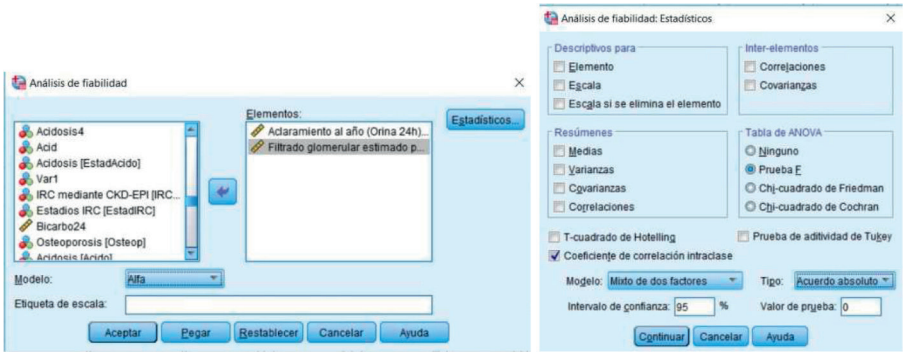
Intraclase (CCI). Vamos a utilizar el Coeficiente de Correlación Intraclase para Acuerdo Absoluto (CCIAb) puesto que es sensible a la presencia de diferencias entre los dos métodos tanto constantes (como el gráfico anterior) como proporcionales (donde las diferencias entre los métodos son diferentes según el valor de las pruebas como podemos ver en el siguiente gráfico:



Vemos que a medida que aumenta el valor de la prueba, las diferencias van siendo proporcionales.

Como ejemplo de esta prueba vamos a utilizar la comparación del filtrado glomerular a través del Aclaramiento de Creatinina al año del trasplante (variable ClCr4 de la base de datos) con el estimado a través de la fórmula CKD-EPI, considerando esta última como método de referencia. Para ello vamos a seguir los siguientes pasos en SPSS:

Analizar → Escalas → Análisis de fiabilidad; se nos abrirá el siguiente cuadro de diálogo:



En el primer cuadro de diálogo, seleccionamos las dos pruebas a comparar y las arrastramos hacia el cuadro "Elementos". Al hacer Clic en el botón "Estadísticos" se nos abre el segundo cuadro de diálogo, donde marcaremos la opción "Prueba F" en la "Tabla de ANOVA", además marcaremos la opción "Coeficiente de correlación Intraclase" y escogeremos la opción "Mixto de dos factores" en "Modelo", y "Acuerdo absoluto" en "Tipo". La sintaxis completa es la siguiente:

```
RELIABILITY
/VARIABLES=ClCr4 CKD_EPI
/SCALE('ALL VARIABLES') ALL
/MODEL=ALPHA /STATISTICS=ANOVA
/ICC=MODEL(MIXED) TYPE(ABSOLUTE) CIN=95 TESTVAL=0.
```

Tras ejecutarla obtenemos el siguiente resultado:

Coeficiente de correlación intraclase

	Correlación intraclase[b]	95% de intervalo de confianza		Prueba F con valor verdadero 0			
		Límite inferior	Límite superior	Valor	gl1	gl2	Sig
Medidas únicas	,438[a]	,031	,685	3,684	60	60	,000
Medidas promedio	,609[c]	,061	,813	3,684	60	60	,000

El CCIAb es de 0.438, con un IC95% entre 0.031 y 0.685, cuyo valor p<0.001. Cuanto más se acerque a 1 el valor de CCIAb lógicamente será mejor pues mejor será la concordancia. En este caso vemos que el valor está bastante lejos de 1, e incluso el límite inferior del IC95% está próximo a 0, luego no parece haber buena concordancia entre ambas formas de medir el filtrado glomerular.

Como norma general existe consenso en valorar la comparación de las pruebas como: Concordancia baja si es <0.40; regular/buena si está entre 0.41 y 0.75, y muy buena si es >0.75.

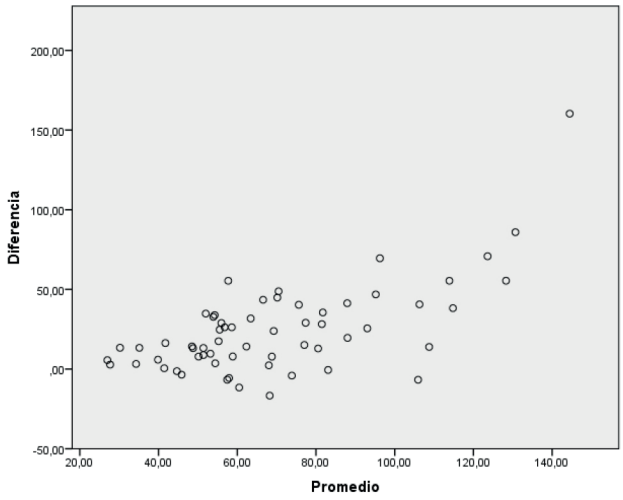
Otra forma de evaluar dos métodos diagnósticos con variables cuantitativas es a través de las diferencias entre los resultados obtenidos por ambas pruebas. Este es el método de Bland-Altman, que aparece con mucha frecuencia en las publicaciones científicas. Consiste en calcular para cada sujeto el valor de la diferencia entre ambas pruebas: di= Xi – Yi, y el valor promedio: p= (Xi + Yi)/2, considerando que este promedio es la mejor estimación posible del verdadero

valor para cada sujeto. Posteriormente se representa de manera gráfica y se interpreta los resultados. Vamos a ver cómo se realiza porque SPSS no tiene implementada esta opción:

En primer lugar debemos crear las variables diferencia y promedio a través de dos COMPUTE. La sintaxis sería:

```
COMPUTE Dif=ClCr4 - CKD_EPI.
COMPUTE Prom=(ClCr4 + CKD_EPI) / 2.
EXECUTE.
```

A continuación representamos de manera gráfica las dos variables obtenidas, eligiendo la variable diferencia como eje Y, y la variable promedio como variable X. El gráfico de puntos obtenido es el siguiente:



Si no existiera diferencia entre ambos métodos, la diferencia entre ambos sería igual a 0, y por tanto todos los puntos se situarían sobre una línea horizontal que pase por el valor 0. Por otra parte si se acepta el supuesto de normalidad (debemos comprobarlo) de los resultados obtenidos con ambas pruebas, es esperable que el 95% de las diferencias se encuentren dentro de los límites que engloban este 95% y que fuera de estos límites se encuentren menos del 0.05% de las diferencias. Para ello debemos calcular la media de las diferencias entre ambas pruebas y los valores de los límites del IC95%. Esta diferencia se obtiene mediante una prueba t-Student para muestras relacionadas. El resultado es el siguiente:

Prueba de muestras emparejadas

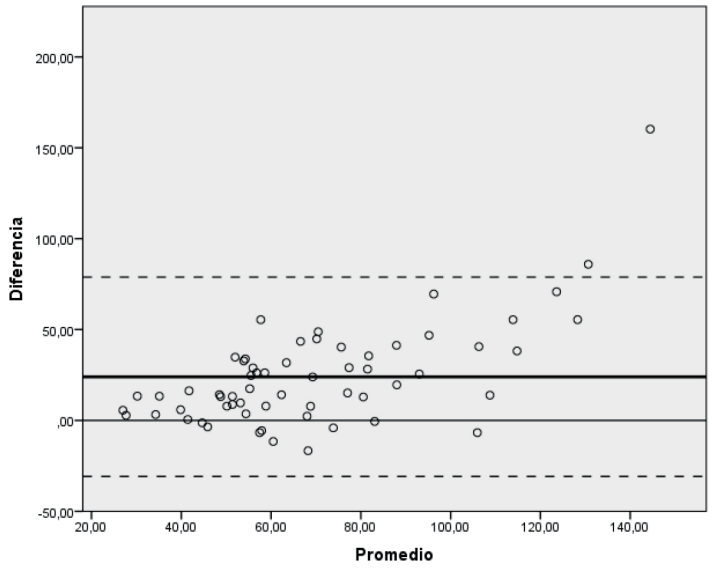
		Diferencias emparejadas					t	gl	Sig. (bilateral)
		Media	Desviación estándar	Media de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior			
Par 1	ClCr4 - CKD_EPI	24,0067	27,9750	3,5818	16,8419	31,1714	6,702	60	,000

La diferencia entre ambas pruebas es de 24.0 ml/min. Los valores de los límites superior e inferior que incluye al 95% de las diferencias es el siguiente:

– Límite inferior: 24.00 – 1.96 x 27.98 = -30.84

– Límite superior: 24.00 + 1.96 x 27.98 = 78.84

Estos valores debemos representarlos en el gráfico anterior haciendo pasar una recta horizontal por cada uno de ellos. Representaremos una recta que pase por el valor 0, por la media de las diferencias y por los valores mínimo y máximo del intervalo del 95% de las diferencias. Para ello debemos seleccionar el gráfico haciendo doble Clic en él y posteriormente doble Clic en uno de los puntos. Elegiremos la opción "Añadir una línea de referencia al eje Y" que aparece en la cinta de opciones de la barra de herramientas; tendremos que hacerlo una vez por cada línea que queramos añadir. Posteriormente podremos editar cada línea para diferenciar unas de otras. El gráfico modificado es el siguiente:



Tras editar el gráfico podemos ver que existe una sobreestimación del filtrado glomerular mediante el aclaramiento de creatinina frente a CKD-EPI (la línea de las diferencias está por encima de la línea del 0). Además podemos ver que las diferencias no son constantes, sino que se modifican a mayor filtrado glomerular (vemos que a partir de 80-90 ml/min de filtrado aproximadamente, las diferencias se hacen mayores). Por último, vemos que tan sólo hay dos puntos fuera del intervalo que incluye al 95% de las diferencias. Para una muestra de 61 sujetos, esperaría encontrar fuera 61 x 0.05 = 3 sujetos; en este caso son 2. Cuantos más puntos estén fuera de este intervalo, menos concordancia hay.

Dado que el procedimiento de Bland-Altman asume que la distribución de la variable diferencia sigue una distribución Normal, habrá que comprobarlo mediante las pruebas de

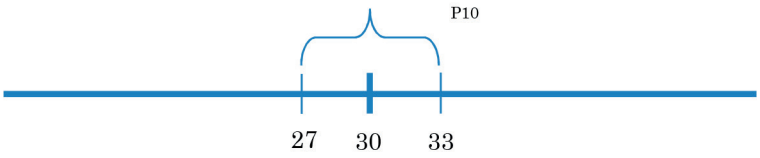
Kolmogorov-Smirnov o Shapiro-Wik según el caso. En este caso el resultado demuestra que no sigue una distribución Normal y por tanto habría que interpretar el gráfico con cautela y analizando los sujetos cuyos valores de la diferencia hacen que no sigan una distribución Normal (corregir si están mal recogidos, eliminarlos si son valores erróneos, etc.).

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Dif	,117	61	,038	,843	61	,000

a. Corrección de significación de Lilliefors

Dentro del estudio de las pruebas diagnósticas con variables cuantitativas, en los últimos tiempos se están definiendo los porcentajes de observaciones de la prueba diagnóstica a validar dentro del intervalo de error del 10, 20 y 30% de la prueba diagnóstica gold standar. ¿Qué quiere decir esto? Supongamos que con la prueba gold standar obtenemos un valor de 30 en la medición de una característica; ¿cuál sería el intervalo que englobaría el 10% de error de esa medición? Pues si de 100 podemos errar en 10, de 30 serían 3; por tanto el intervalo de error al 10% estaría situado entre 27 y 33, como vemos en el siguiente gráfico:



Representa la precisión y exactitud de la nueva prueba diagnóstica a validar. El intervalo anterior entre 27 y 33 podemos considerarlo el centro de una diana y lo que nos interesa es ver cuantas observaciones de la nueva prueba diagnóstica se encuentran dentro de ese intervalo que sería el centro de la diana. Cuantas más haya, más exactitud y precisión tendrá la prueba. Se suelen utilizar los intervalos al 10%, 20% y 30%. SPSS esto no lo calcula y tenemos

que hacerlo "a mano". Basta con crear nuevas variables denominadas P10, P20 y P30 con la fórmula siguiente:

$$P = \left|\frac{GE - Nu}{GE}\right| \times 100$$

Dónde:

– GE: Prueba gold estándar.

– Nu: Nueva prueba diagnóstica.

– P: porcentaje de error.

En el resultado de la fracción tomaremos el valor absoluto pues puede darnos valores negativos según la dirección de la diferencia del numerado.

Como ejemplo vamos a ver el porcentaje de observaciones de ClCr que se encuentran dentro del 10% de error de CKD-EPI. En primer lugar creamos la variable P con la siguiente sintaxis:

```
COMPUTE P=(ABS((CKD_EPI - ClCr4)/CKD_EPI)) * 100.
EXECUTE.
```

De esta forma se nos crea la variable P con el porcentaje de errores para cada caso. Ahora tenemos que ver qué porcentaje de esos es inferior a 10 mediante otro COMPUTE:

```
COMPUTE P10=P < 10.
EXECUTE.
```

Así se nos creará una variable binaria con los valores correspondientes a un error <10% etiquetados como 1.

Por último utilizaremos el procedimiento FRECUENCIES para ver el "porcentaje de 1" etiquetados como "Sí" al definir la variable:

```
FREQUENCIES VARIABLES=P10
/ORDER=ANALYSIS.
```

Cuyo resultado es:

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	No	51	69,9	83,6	83,6
	Sí	10	13,7	16,4	100,0
	Total	61	83,6	100,0	
Perdidos	Sistema	12	16,4		
Total		73	100,0		

En este caso vemos que tan sólo el 16.4% de las observaciones se encuentran dentro del intervalo del 10% de error de CKD-EPI. Podemos ser más "finos" aún y describir además el IC95% de esta observación. Se realiza calculando el IC95% de una proporción, cuya fórmula es:

$$IC95\% = P0 \pm 1.96 \times \sqrt{\frac{Po(1-P0)}{n}}$$

Dónde:

– P_0: probabilidad observada.

– N: tamaño de la muestra.

Para el ejemplo anterior sería:

$$IC95\% = 0.164 \pm 1.96 \times \sqrt{\frac{0.164 \times 0.836}{61}} = 0.164 \pm 0.0929$$

El IC95% en este caso por tanto se sitúa entre: 0.0711 y 0.2569; 7.1 y 25.7.

Por tanto en el ejemplo diremos que el porcentaje P10% de error es de 16.4% IC95%(7.1% y 25.7%).

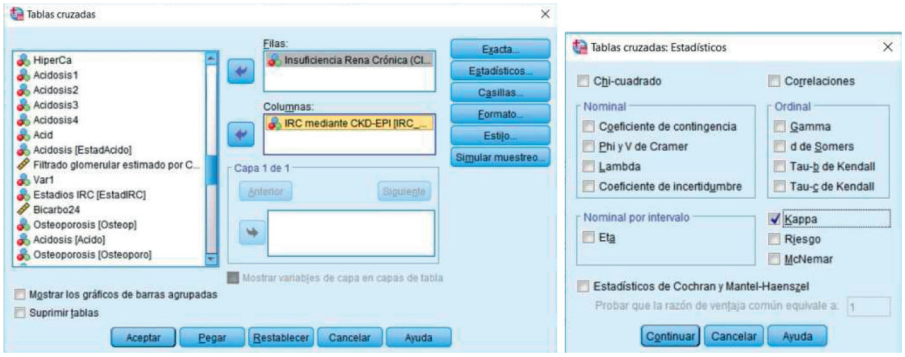
Con P20% y P30% se procede de la misma manera.

ANÁLISIS ESTADÍSTICOS PARA PRUEBAS DIAGNÓSTICAS CON VARIABLES CATEGÓRICAS

En este caso el resultado de la prueba diagnóstica será una variable categórica con 2 o más categorías. La comparación entre ellas se realiza mediante el índice Kappa. Evalúa las frecuencias observadas frente a las esperadas en cada casilla si el acuerdo fuera absoluto.

Como ejemplo vamos a comparar la estratificación en pacientes con o sin IRC (FG<60 ml/min) mediante el filtrado glomerular estimado por aclaramiento de creatinina (ClCr) y mediante la fórmula CKD-EPI. En la base de datos hemos creado las variables "IRC" donde los pacientes se dividen en: 0 sin IRC y 1 con IRC a través del ClCr; y la variable "IRC-CKD" donde los pacientes se dividen de igual manera pero esta vez a través de CKD-EPI. Vamos a comparar la asignación de pacientes a los grupos con y sin IRC por ambos métodos. Para realizarlo vamos a seguir el siguiente procedimiento:

Analizar → Estadísticos descriptivos → Tablas cruzadas; se nos abrirá el cuadro de diálogo habitual para las tablas de contingencia de 2x2:



Seleccionamos las dos variables a comparar poniendo una en "Filas" y la otra en "Columnas". Hacemos Clic en el botón "Estadísticos" y se nos abrirá el segundo cuadro de diálogo donde vamos a marcar la opción "Kappa". La sintaxis es la siguiente:

```
CROSSTABS
/TABLES=IRC BY IRC_CKD
/FORMAT=AVALUE TABLES
/STATISTICS=KAPPA
/CELLS=COUNT
/COUNT ROUND CELL.
```

Tras ejecutarla obtenemos los siguientes resultados:

Tabla cruzada IRC*IRC_CKD

Recuento

		IRC_CKD		
		No	Sí	Total
CICr	No	23	17	40
	Sí	4	17	21
Total		27	34	61

Medidas simétricas

		Valor	Error estandarizado asintótico[a]	T aproximada[b]	Significación aproximada
Medida de acuerdo	Kappa	,335	,108	2,873	,004
N de casos válidos		61			

a. No se presupone la hipótesis nula.

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

En la primera tabla se representa la distribución de los pacientes en una u otra categoría por ambos métodos. Los pacientes que están distribuidos en la misma categoría por ambos métodos son los que se encuentran en la diagonal: hay 23 pacientes etiquetados como sin IRC y 17 pacientes etiquetados como con IRC por ambos métodos. En estos pacientes ambos

métodos coinciden. Cuanto mayor sea el número de sujetos en esta diagonal, mayor será la concordancia entre ambas pruebas.

El índice Kappa evalúa el porcentaje de sujetos que se encuentran dentro de la diagonal una vez extraído el porcentaje de sujetos que están bien etiquetados simplemente por azar. Su valor oscila entre -1 (desacuerdo absoluto) y 1 (acuerdo absoluto). Un valor de Kappa=0 indica asociación debida al azar.

Existe consenso para establecer que la concordancia entre dos métodos evaluadas por el índice Kappa es pobre cuando es <0.40, moderada entre 0.40 y 0.75, y fuerte entre 0.76 y 1.00. En este caso el valor es de 0.34, por lo tanto la concordancia entre ClCr y CKD- EPI para etiquetar un paciente con o sin IRC es pobre. Su IC95% se calcula a través del error estándar que aparece en la tabla:

– Límite inferior: 0.335 - 1.96 x 0.108 = 0.123.

– Límite superior: 0.335 + 1.96 x 0.108 = 0.547.

En este caso hemos comparado dos pruebas diagnósticas con variables categóricas binarias. Cuando las variables tienen más de 2 categorías el índice Kappa no es apropiado porque no tiene en cuenta si las discrepancias son más importantes en unas categorías que en otras, las trata todas por igual. Por otra parte, simplemente por el hecho de aumentar el número de categorías se produce un aumento del valor de Kappa sin que ello signifique que hay más acuerdo. Para ello se debe hacer el índice Kappa Ponderado que asigna una ponderación (w) diferente a cada categoría, siendo la ponderación más alta en los extremos (las más "castigadas" puesto que son las más alejadas de la diagonal y por tanto las menos concordantes). Pero SPSS no tiene implementado esta opción. Para ello lo que debemos hacer es realizar agrupaciones de categorías y compararlas entre sí, o bien realizar el cálculo "a mano". Para ello, en primer lugar le daremos la ponderación correspondiente a cada casilla, de manera que las casillas sobre la diagonal la ponderación será 0 (puesto que son las realmente concordantes), mientras que a medida que nos alejemos de la diagonal le otorgaremos como ponderación el valor de la distancia a la diagonal al cuadrado. Así para una tabla de 4x4 el valor de las ponderaciones sería:

Método A	Método B			
		$W=1^2=1$	$W=2^2=4$	$W=3^2=9$
	$W=1^2=1$		$W=1^2=1$	$W=2^2=4$
	$W=2^2=4$	$W=1^2=1$		$W=1^2=1$
	$W=3^2=9$	$W=2^2=4$	$W=1^2=1$	

La fórmula para el cálculo del Índice Kappa ponderado es la siguiente:

$$IKp = 1 - \frac{\sum Wij \times Oij}{\sum Wij \times eij}$$

Dónde:

– W_{ij}: es la ponderación correspondiente a la casilla.

– o_{ij}: Corresponde a la frecuencia observada de cada casilla.

– e_{ij}: Corresponde a la frecuencia esperada en cada casilla.

Como ejemplo vamos a calcular el valor de IKp utilizando como prueba diagnóstica el estadio de función renal (Estadio de IRC) obtenido por CKD-EPI y ClCr. En primer lugar realizaremos la tabla con las frecuencias observadas y esperadas de cada casilla.

El resultado es:

Tabla cruzada EstadIRCClCr*EstadIRC

			EstadIRC			
			Sin IRC	Estadio 3	Estadio 4	Total
EstadIRCClCr	Sin IRC	Recuento	23	17	0	40
		Recuento esperado	17,7	19,7	2,6	40,0
	Estadio 3	Recuento	4	13	2	19
		Recuento esperado	8,4	9,3	1,2	19,0
	Estadio 4	Recuento	0	0	2	2
		Recuento esperado	,9	1,0	,1	2,0
Total		Recuento	27	30	4	61
		Recuento esperado	27,0	30,0	4,0	61,0

El valor del IKp sería el siguiente:

$$IKp = 1 - \frac{(1 \times 17) + (4 \times 0) + (1 \times 4) + (1 \times 2) + (4 \times 0) + (1 \times 0)}{(1 \times 19.7) + (4 \times 2.6) + (1 \times 8.4) + (1 \times 1.2) + (4 \times 0.9) + (1 \times 1.0)} = 0.519$$

Este valor es discretamente superior al obtenido sin la ponderación, cuyo resultado es 0.320.

Por último hay que decir que no todo el mundo está de acuerdo con la utilización del índice Kappa como medidor de la concordancia entre pruebas diagnósticas, aunque es el procedimiento actual utilizado y el que aparece en todas las publicaciones.

CONCEPTOS DE SENSIBILIDAD Y ESPECIFICIDAD EN PRUEBAS DIAGNÓSTICAS

El objetivo final de cualquier prueba diagnóstica lógicamente es determinar a través de la prueba si un sujeto presenta o no una determinada característica. Por ejemplo, decidir si un paciente presenta o no Lupus en función de la positividad de los anticuerpos antiDNA.

Para validar cualquier prueba diagnóstica en su fase de experimentación se tiene que demostrar que es una prueba fiable, estableciéndose durante esta fase los conceptos de Sensibilidad y Especificidad que se interpretan del siguiente modo:

- Sensibilidad (Se): Capacidad que tiene una prueba diagnóstica de ser positiva en los sujetos que presentan una característica. Por ejemplo, que el antiDNA sea positivo en los pacientes con Lupus.

- Especificidad (Sp): Capacidad que tiene una prueba diagnóstica para ser negativa en los sujetos que no presentan dicha característica. Por ejemplo, negatividad del antiDNA en los pacientes sin Lupus.

De las anteriores definiciones surgen también los conceptos de falsos positivos y falsos negativos:

- Falso positivo (FP): positividad de la prueba diagnóstica en sujetos sin la característica. Por ejemplo, positividad del antiDNA en sujetos sin Lupus.
- Falso negativo (FN): negatividad de la prueba diagnóstica en sujetos con la característica. Por ejemplo, negatividad del antiDNA en sujetos con Lupus.

Los sujetos con una prueba positiva y que realmente presentan la característica serán los Verdaderos positivos (VP), mientras que los sujetos con la prueba negativa que no tienen la característica será los Verdaderos negativos.

Así la Sensibilidad se calcula del siguiente modo:

$$Se = \frac{VP}{VP + FN}$$

Y para la Especificidad sería:

$$Sp = \frac{VN}{VN + FP}$$

Lo ideal para una prueba diagnóstica es que la sensibilidad y la especificidad sean lo mayor posible, pero sus valores se modifican de manera inversa, de manera que a mayor Se menor Sp. Llegar a establecer un punto de corte que magnifiquen en la medida de lo posible ambas pruebas es uno de los objetivos de los estudios que validan pruebas diagnósticas. Por ejemplo, establecer el punto de corte de la determinación de la PCR para virus CMV para decidir si un paciente presenta o no enfermedad por CMV.

Para que una prueba diagnóstica sea aceptable, los valores de Se y Sp deben estar lo más próximo a 1 posible. Se recomienda que al menos sus valores sean mayores a 0.75.

Pero una vez que una prueba diagnóstica se ha validado para su uso, debemos determinar si una persona presenta o no una característica en función de si la prueba es positiva o negativa. Es decir, es el procedimiento inverso al que se llevó a cabo durante la fase de experimentación de la prueba. De esto surgen los conceptos de Valor Predictivo Positivo (VPP) y Valor Predictivo Negativo (VPN), que se interpretan de la siguiente manera:

- VPP: Probabilidad que tiene un sujeto de presentar una característica cuando la prueba es positiva. Por ejemplo, probabilidad de que un paciente tenga Lupus cuando los antiDNA son positivos.

- VPN: Probabilidad que tiene un sujeto de no presentar una característica cuando la prueba es negativa. Por ejemplo, probabilidad que tiene un sujeto de no tener Lupus cuando los antiDNA son negativos.

En la práctica diaria habitual son estos dos conceptos los que nos interesan pues son los que van a condicionar nuestra actitud ante un sujeto.

Las fórmulas para su cálculo son:

$$VPP = \frac{VP}{VP + FP}$$

$$VPN = \frac{VN}{VN + FN}$$

Vamos a hacer los cálculos tomando como ejemplo nuestra base de datos. Para ello vamos a suponer que CKD-EPI es la forma más exacta de estimar el FG de un paciente y que cataloga de manera correcta a un paciente con o sin IRC, y que el aclaramiento de creatinina ClCr es la nueva prueba diagnóstica que queremos validar. Tomamos los datos de la tabla de contingencia del apartado anterior:

Tabla cruzada IRC_CKD*IRC

Recuento

		ClCr		
		No	Sí	Total
IRC_CKD	No	23	4	27
	Sí	17	17	34
Total		40	21	61

- 17 pacientes presentan la prueba positiva en ambos métodos: VP= 17.
- 23 pacientes presentan la prueba negativa con ambos métodos: VN=23.
- 17 pacientes tienen prueba negativa con ClCr cuando realmente sí tienen IRC por CKD: FN=17.
- 4 pacientes presentan la prueba positiva cuando realmente no tienen IRC por CKD: FP=4.

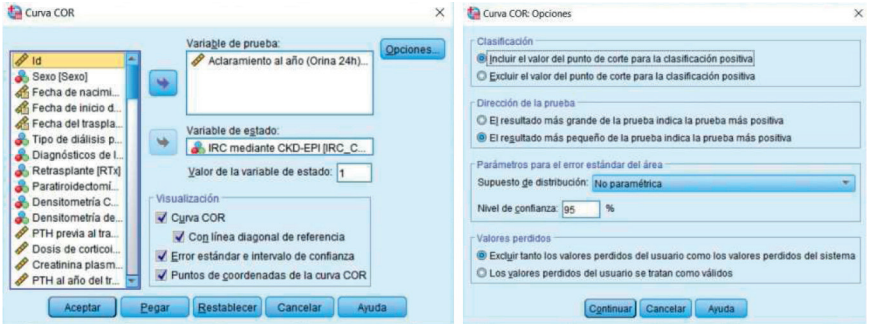
A partir de estos datos podemos calcular los resultados de los conceptos anteriores:

- Se= 17/(17+17)= 0.50.
- Sp= 23/(23 + 4)= 0.85.
- VPP= 17/(17+4)= 0.81.
- VPN= 23/(23+17)= 0.58

La aplicación de una prueba diagnóstica en población general para detectar una determinada característica va a depender del objetivo de actuación. Por ejemplo, si se quiere detectar la mayor cantidad posible de sujetos que presentan un tumor en la población general asintomática se debe utilizar una prueba lo más sensible posible para que no quede ningún sujeto sin identificar; pero vamos a necesitar después otra prueba que sea más específica que nos descarte los falsos positivos. Así sucede por ejemplo con el PSA para detectar tumor prostático: la sensibilidad es alta y por tanto detectará a la mayoría de sujetos con tumor, pero su especificidad es baja y por tanto habrá muchos falsos positivos, que habrá que descartar posteriormente con otra prueba con mayor especificidad, como una biopsia prostática, la cual es más cara y también más cruenta. Por tanto, las pruebas de cribaje en población general se suelen realizar con métodos diagnósticos con alta sensibilidad, baratos y poco cruentos (PSA, mamografía, etc.).

De manera gráfica podemos representar la relación entre Se y Sp de una prueba diagnóstica. Esto se hacer mediante curvas ROC (receiver operating characteristic, sin traducción al español). Este procedimiento además nos permite establecer un punto de corte a partir del cual establecer un diagnóstico. Por ejemplo, imaginemos que estamos realizando un estudio donde queremos establecer un punto de corte del ClCr que nos determine que un paciente presenta IRC con bastantes garantías (determinado el FG mediante un método fiable que en este caso es CKD-EPI). Para ello necesitamos que tanto su Se como su Sp sean las mejores posibles. Para ello debemos calcular la Se y la Sp para cada valor del ClCr de la muestra. Mediante SPSS se realiza de la siguiente manera:

Analizar → Curva COR; se nos abrirá el siguiente cuadro de diálogo:



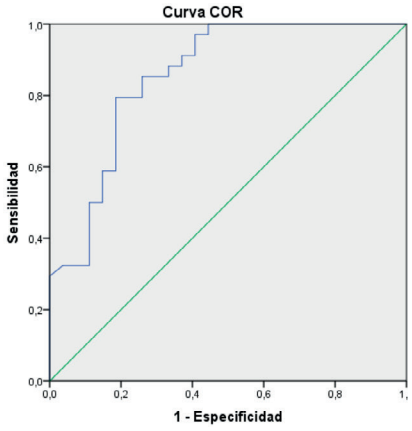
En el primer cuadro de diálogo vamos a seleccionar la variable con la prueba diagnóstica a estudio, en este caso "ClCr" y la arrastramos hacia el cuadro "Variable de prueba". En el cuadro "Variable de estado" vamos a seleccionar la variable con el método diagnóstico de referencia, en este caso "CKD-EPI". En el cuadro de "Valor de la variable estado" debemos introducir el valor de la categoría que representa la característica a estudio, en este caso el valor 1 que es el que representa a los pacientes con IRC en la muestra. Marcamos las opciones "Curva COR", "Con línea diagonal de referencia", "Error estándar e intervalo de confianza" y "Puntos de coordenada de la curva COR". Después hacemos Clic en el

botón "Opciones" y en este ejemplo debemos seleccionar la opción "El resultado más pequeño de la prueba indica la prueba más positiva" ya que los valores más pequeños de ClCr son los que van a determinar que un paciente tenga IRC (prueba más positiva porque CKD-EPI con el valor 1 son los que tienen IRC y con el valor 0 los que no). En otras circunstancias habría que seleccionar la opción "El resultado más grande de la prueba indica la prueba más positiva" que es la seleccionada por defecto; por ejemplo si queremos evaluar que las cifras de tensión arterial determinan que un paciente sea hipertenso: a mayor cifra de tensión arterial, mayor probabilidad de ser hipertenso.

La sintaxis es la siguiente:

```
ROC ClCr4 BY IRC_CKD (1)
/PLOT=CURVE(REFERENCE)
/PRINT=SE COORDINATES
/CRITERIA=CUTOFF(INCLUDE) TESTPOS(SMALL) DISTRIBUTION(FREE) CI(95)
/MISSING=EXCLUDE.
```

Tras ejecutarla obtenemos los siguientes resultados:



Área bajo la curva

Variables de resultado de prueba: ClCr4

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,856	,050	,000	,758	,953

Las variables de resultado de prueba: ClCr4 tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

En primer lugar se representa el gráfico con los valores de la Se en el eje de ordenadas X y la Sp en el eje de abscisas Y (aunque aparece su complementario, es decir, 1-S). Cuanto más hacia arriba y hacia la izquierda mejor. Un gráfico que se sitúe próximo a la diagonal carece que capacidad diagnóstica y su área bajo la curva (AUC) valdrá 0.5. Cuanto más se aproxime a 1 el AUC mayor capacidad diagnóstica. En este caso el AUC vale 0.856, con un IC95% entre 0.758 y 0.953, con una $p<0.001$, que podemos considerar como bastante aceptable.

A continuación aparece una tabla con los valores de la Se y la Sp (en forma de 1-Sp) para cada valor del ClCr:

Coordenadas de la curva

Variables de resultado de prueba: ClCr4

Positivo si es menor o igual que[a]	Sensibilidad	1 - Especificidad
28,10	,000	,000
29,45	,029	,000
32,85	,059	,000
36,40	,088	,000
39,30	,118	,000
41,75	,147	,000
42,30	,176	,000
43,40	,206	,000
44,05	,235	,000
47,00	,265	,000
52,00	,294	,000
54,40	,324	,037
54,90	,324	,074
55,20	,324	,111
55,45	,353	,111
55,70	,382	,111
56,00	,412	,111
57,10	,441	,111
58,95	,500	,111
61,35	,500	,148
63,40	,529	,148
65,95	,559	,148
68,55	,588	,148
69,30	,588	,185
69,65	,647	,185
70,10	,676	,185
70,35	,706	,185
70,80	,735	,185
71,45	,765	,185
71,75	,794	,185
72,25	,794	,222
76,00	,794	,259
80,25	,824	,259
82,00	,853	,259
83,70	,853	,296
85,00	,853	,333

En esta tabla vamos a buscar el valor de la prueba (ClCr) que mayor Se y Sp tienen. Debemos tener en cuenta que la Sp se representa como 1-Sp. Si hemos dicho que la Sp debe ser al menos de 0.75, su complementario (1-Sp) será 0.25. Es decir, vamos a buscar en la tabla aquel valor cuya Se sea mayor a 0.75 y Sp menor a 0.25. Si observamos la tabla, vemos que el valor de ClCr igual a 71.75 es el que mayor Se y Sp presentan ya que la Se=0.794 y la Sp=0.815. Por tanto podríamos decir que pacientes con un valor de ClCr inferiores a 71.75 ml/min, tienen alta probabilidad de presentar IRC. Si creamos una nueva variable que represente la presencia de IRC por ClCr con el punto de corte 71.75 ml/min, y calculamos el nuevo valor Kappa, veremos que la concordancia entre la nueva variable creada y IRC-CKD ha aumentado.

UNIVERSIDAD DE EXTREMADURA